**UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ**
**DEPARTAMENTO ACADÊMICO DE INFORMÁTICA**
**CURSO DE ENGENHARIA DE COMPUTAÇÃO**

**ELZA MEIRA PUPPO**

# SISTEMA DE RECOMENDAÇÃO DE INSTITUIÇÃO DE ENSINO SUPERIOR PARA ATENUAR EVASÃO UTILIZANDO ANÁLISE DE PERFIS DE EGRESSOS

**TRABALHO DE CONCLUSÃO DE CURSO**

**PATO BRANCO**
**2021**

**ELZA MEIRA PUPPO**

# SISTEMA DE RECOMENDAÇÃO DE INSTITUIÇÃO DE ENSINO SUPERIOR PARA ATENUAR EVASÃO UTILIZANDO ANÁLISE DE PERFIS DE EGRESSOS

Trabalho de Conclusão de Curso apresentado à disciplina de Trabalho de Conclusão de Curso 2, do Curso de Engenharia de Computação da Universidade Tecnológica Federal do Paraná, Câmpus Pato Branco, como requisito parcial para obtenção do título de Bacharel em Engenharia de Computação.

Orientador: Prof. Dr. Ives Rene Venturini Pola

**PATO BRANCO**

**2021**

**Ministério da Educação**
**Universidade Tecnológica Federal do Paraná**
**Campus Pato Branco**
**Curso de Engenharia de Computação**

UTFPR
UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

## TERMO DE APROVAÇÃO

Às 8 horas e 30 minutos do dia 19 de maio de 2021, reuniu-se de forma online a banca examinadora composta pelos professores Prof. Dr. Ives Renê Venturini Pola (orientador), Profa. Dra. Viviane Dal Molin de Souza e Profa. Dra. Mariza Miola Dosciatti para avaliar o trabalho de conclusão de curso com o título **Sistema de recomendação de instituição de ensino superior para atenuar evasão utilizando análise de perfis de egressos**, da aluna **Elza Meira Puppo**, matrícula 1271237, do curso de Engenharia de Computação. Após a apresentação o aluno foi arguido pela banca examinadora. Em seguida foi realizada a deliberação pela banca examinadora que considerou o trabalho aprovado.

_____
Prof. Dr. Ives Renê Venturini Pola
Orientador (UTFPR)

_____
Profa. Dra. Viviane Dal Molin de Souza
(UTFPR)

_____
Profa. Dra. Mariza Miola Dosciatti
(UTFPR)

_____
Prof. Dr. Pablo Gauterio Cavalcanti
Coordenador de TCC

_____
Prof. Dr. Ives Renê Venturini Pola
Coordenador do Curso de
Engenharia de Computação

*Experience is merely the name men gave to*

*their mistakes.*

The Picture of Dorian Gray - Oscar Wilde

## RESUMO

A evasão no ensino superior é um desperdício social, acadêmico e econômico. Estudos têm sido realizados de maneira a identificar os fatores que levam à evasão e, da mesma forma, como combater tais fatores. Este trabalho propõe a utilização de métodos para Extração de Conhecimento e Sistema de Recomendação com a finalidade de sugerir instituições de ensino a futuros ingressantes. O primeiro processo realizado consistiu em compreender a base do Exame Nacional de Desempenho dos Estudantes (ENADE), limpar, pré-processar, transformar os dados e realizar tarefas de Mineração de Dados para obter perfis dos egressos. As tarefas de Mineração de Dados utilizadas abrangeram técnicas dos métodos de agrupamento e classificação para definir os perfis. O segundo processo executado foi o desenvolvimento de um Sistema de Recomendação, por meio de uma aplicação *Web* para a coleta dos dados do usuário a fim de gerar um perfil desse. Nesse sistema, o perfil do usuário é comparado com os perfis extraídos no primeiro processo a fim de identificar a similaridade entre eles. Posteriormente, o sistema recomenda as instituições de ensino que possuem um perfil mais similar ao perfil do usuário.



**Palavras-chave:** Análise por agrupamento. Evasão universitária. Descoberta de conhecimento em base de dados. Mineração de dados.

## ABSTRACT

Evasion in higher education is a social, academic and economic waste. Studies have been carried out in order to identify the coefficients that lead to evasion and, likewise, verify how to oppose such coefficients. This work proposes the use of methods for Knowledge-Discovery in Databases and Recommendation System in order to suggest educational institutions to future students. The first process performed consisted of understanding the database of ENADE, cleaning, pre-processing, transforming the data and performing Data Mining tasks to obtain profiles of the graduates. The Data Mining tasks used covered techniques of clustering and classification methods to define the profiles. The second process performed was the development of a Recommendation System, using a Web application to collect user data in order to generate a profile. In this system, the user's profile is compared with the profiles extracted in the first process in order to identify the similarity between them. Subsequently, the system recommends educational institutions that have a profile more similar to the user's profile.



**Keywords:** Clustering analysis. University dropouts. Knowledge discovery in databases. Data mining.

## LISTA DE ILUSTRAÇÕES

## LISTA DE TABELAS

## LISTA DE SIGLAS E ACRÔNIMOS

SIGLAS

| | |
|---|---|
| CSS | *Cascading Style Sheets* |
| CSV | *Comma-Separated Values* |
| HTML | Linguagem de Marcação de Hipertexto |
| PHP | *Hypertext Preprocessor* |

ACRÔNIMOS

| | |
|---|---|
| DBSCAN | *Density-Based Clustering Based on Connected Regions with High Density* |
| DENCLUE | *Clustering Based on Density Distribution Functions* |
| ENADE | Exame Nacional de Desempenho dos Estudantes |
| ENEM | Exame Nacional do Ensino Médio |
| INEP | Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira |
| MEC | Ministério da Educação |
| OPTICS | *Ordering Points to Identify the Clustering Structure* |
| ProUni | Programa Universidade para Todos |
| SiSU | Sistema de Seleção Unificada |

**SUMÁRIO**

# 1 INTRODUÇÃO

O ingresso e a permanência de alunos no ensino superior no Brasil ainda não é uma realidade para todos. Mesmo com a criação de programas de avaliação e de democratização do acesso a cursos de graduação, como o Exame Nacional do Ensino Médio (ENEM), o Sistema de Seleção Unificada (SiSU) e o Programa Universidade para Todos (ProUni), em 2017, segundo PORTAL INEP (2018), o número de vagas remanescentes chegou a 2,8 milhões. Além disso, de acordo com Dados do Censo da Educação Superior (2016), a taxa de evasão no ensino superior atingiu 49% em 2014.

A evasão é um desperdício social, acadêmico e econômico em virtude da ociosidade de professores, funcionários, equipamentos e espaço físico. Deste modo, a evasão no setor público representa investimentos sem o devido retorno, e no setor privado, a redução de receitas (FILHO *et al.*, 2007). Sendo assim, é necessário analisar suas causas e incluir programas institucionais com o propósito de reduzir esses índices (DAVOK; BERNARD, 2016).

Esse tema tem sido abordado em diversas pesquisas que descrevem as causas da evasão competentes a cada indivíduo e competentes à comunidade acadêmica da instituição de ensino (BIAZUS *et al.*, 2004). Paredes (1994), Martins (2007) e Scali (2009) retratam características particulares do perfil do ingressante como, por exemplo, a necessidade de realizar trabalho remunerado durante a graduação e não ser capaz de equilibrar os estudos e o trabalho. No entanto, Tinto (1987) aponta outros fatores que influenciam na evasão do acadêmico como o histórico familiar; o conjunto de relações formais e informais estabelecido no ambiente acadêmico e social, como performance acadêmica e atividades extra-curriculares; e, por fim, a integração acadêmica e social que os itens anteriores proporcionam. Algumas instituições de ensino buscam soluções que intentam amenizar esses fatores.

Em geral, essas soluções proporcionam maior interação entre professores e alunos, melhorias das condições de infraestrutura, acompanhamento e atendimento aos estudantes, incentivo às habilidades, respeito às individualidades, além do apoio financeiro e psicológico (OLIVEIRA *et al.*, 2019). Por exemplo, Machado *et al.* (2005) retratam o incentivo às bolsas de iniciação científica e a possibilidade de mudança de turno do curso, disponibilizando horário de expediente aos acadêmicos que precisam trabalhar. Almeida *et al.* (2019) descrevem as atividades de incentivo à adaptação acadêmica por meio de oficinas com instruções financeiras, profissionais, emocionais e pedagógicas. Além disso, Silva *et al.* (2019) avaliam ações pré-universitárias como

seminários ou palestras sobre os cursos e orientação vocacional, bem como a oferta de monitorias ou atendimento extraclasse para suprir deficiências do aluno ao longo do curso. Por fim, Rigo *et al.* (2014) mostram que o grande volume de dados na área de educação tem fomentado o interesse da aplicação de técnicas de Extração de Conhecimento para dar suporte a práticas pedagógicas, especificamente no contexto deste trabalho, técnicas para redução de evasão.

Extração de Conhecimento é a aquisição de informação implícita, anteriormente desconhecida, e potencialmente útil de dados (WITTEN; FRANK, 2002). Ela é amplamente utilizada em áreas como negócios (GIUDICI; FIGINI, 2009), segurança (BARBARÁ; JAJODIA, 2002), finanças (NGAI *et al.*, 2011), medicina (LAVRAČ, 1999), agricultura (MUCHERINO *et al.*, 2009), entre outras. A Mineração de Dados Educacionais é a aplicação de técnicas de Mineração de Dados, parte do processo de Extração de Conhecimento, em dados educacionais para entender melhor os estudantes, seu cenário de aprendizado, efetuar acompanhamento e previsões (ROMERO; VENTURA, 2010). A Mineração de Dados Educacionais pode ser utilizada com diversas finalidades como, por exemplo, melhorar as metodologias de ensino, melhorar o desempenho dos acadêmico, avaliar o uso de plataformas de ensino virtuais, rastrear causas de evasão escolar, entre outras. Dentre os atuais métodos de Mineração de Dados Educacionais com abordagem de evasão estudantil, destacam-se as tarefas de agrupamento e os sistemas de recomendação.

As tarefas de agrupamento visam identificar e aproximar os registros similares entre si, distanciando-os dos registros dos demais agrupamentos (CAMILO; SILVA, 2009). O resultado dessa tarefa permite a análise dos resultados que, inserida no domínio do contexto, favorece a compreensão e reflexão capazes de validar hipóteses e construir conhecimento. Já o método de sistemas de recomendação, segundo Sarwar *et al.* (2000), é dividido em quatro processos: identificação do usuário, coleta de informações, estratégias de recomendação e visualização das recomendações, tendo como propósito a recomendação de produto, serviço ou conteúdo de acordo com as necessidades e interesses do usuário.

No contexto da Mineração de Dados Educacionais utilizando um método de agrupamento, citam-se Paz e Cazella (2017), que realizaram um estudo de caso que buscou identificar perfis de alunos com potencial de evasão por meio da aplicação de técnicas de Extração do Conhecimento. Da mesma forma, Li *et al.* (2011) propõem uma abordagem que descobre automaticamente perfis de alunos usando um agente de aprendizagem de máquina, criado a partir dos padrões de comportamento dos alunos.

Durand *et al.* (2011) apresentam um sistema de recomendação de planos de ensino,

enquanto que Toscher e Jahrer (2010) utilizam técnicas de filtragem colaborativa para desenvolver um sistema de recomendação com base nos históricos de acadêmicos. Jiménez-Raygoza *et al.* (2019) propõem um *framework* que utiliza base de dados alimentada por meio de formulários on-line para ajudar os alunos a determinar as possibilidades de sucesso na escolha de uma carreira universitária. Já Hasebrook e Nathusius (1997) desenvolveram um sistema que combina uma enciclopédia profissional e testes vocacionais para auxiliar nas decisões de carreira adequadas, simulando a avaliação de um especialista.

Esses estudos, no geral, visam evitar a evasão no ensino superior considerando os acadêmicos já matriculados no curso para identificar as causas, esboçar perfis que revelam risco de abandono do curso, ou indicar possíveis falhas nas técnicas de ensino. Contudo, pelo que é de conhecimento dos autores deste documento, não há estudos que utilizem as características individuais dos participantes do ENADE nos anos retroativos para conceber perfis que possam ser utilizados para recomendação a futuros acadêmicos.

O ENADE é uma prova realizada anualmente pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) com a finalidade de avaliar: o rendimento dos concluintes dos cursos de graduação em relação aos conteúdos programáticos previstos nas diretrizes curriculares dos cursos; o desenvolvimento de competências e habilidades necessárias ao aprofundamento da formação geral e profissional; e o nível de atualização dos estudantes com relação à realidade brasileira e mundial (INEP, 2020a). Além dos componentes de rendimento, o ENADE apresenta o Questionário do Estudante que tem como objetivo levantar informações que permitam caracterizar o perfil dos estudantes e o contexto de seus processos formativos (INEP, 2020c).

O Questionário do Estudante possui perguntas como: O acadêmico realiza trabalho remunerado enquanto estuda? Participou de atividades curriculares no exterior? Teve oportunidade de aprender um outro idioma na instituição? Qual o nível de escolaridade dos pais? Quantas horas semanais foram dedicadas ao estudo extraclasse? Além disso, contém uma seção em que o acadêmico avalia a instituição de ensino em quesitos como: organização didático-pedagógica; organização de infraestrutura; instalações físicas; e oportunidade de ampliação da formação acadêmica e profissional. Essas informações, providas pelo INEP, podem ser comparadas a fim de verificar a influência que possuem na escolha do curso e no desempenho do acadêmico.

## 1.1 OBJETIVOS

Desenvolver um sistema *Web* de recomendação para ingressantes de instituições de ensino superior com base em perfis de egressos. Nesse sistema o aluno escolhe o curso de interesse e o sistema sugere a instituição de ensino.

Para atingir tal objetivo, executa-se as seguintes tarefas:

- Compreender e selecionar os dados extraídos dos portais oficiais do governo brasileiro;
- Pré-processar as bases, realizando tratamento de dados faltantes e correção de anomalias;
- Transformar os dados, selecionando, normalizando e codificando os atributos relevantes a fim de torná-los compatíveis com a etapa posterior;
- Aplicar tarefas de Mineração de Dados com base nos perfis dos participantes do ENADE;
- Interpretar e avaliar o conhecimento descoberto; e
- Desenvolver uma aplicação *Web* com sistema de recomendação e interface para usuário.

# 2 REFERENCIAL TEÓRICO

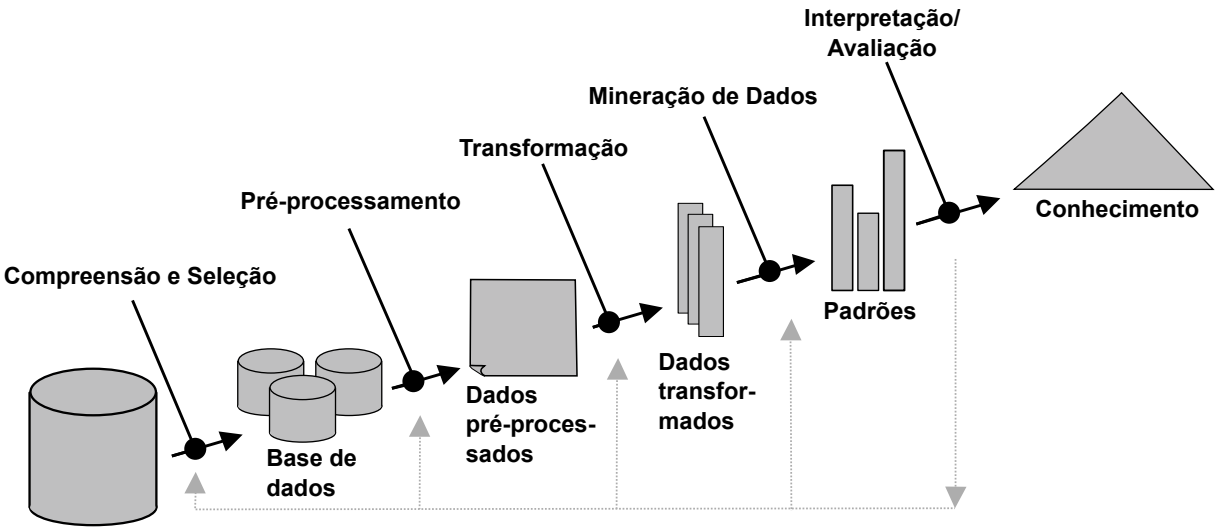
Com a evolução dos recursos computacionais, informatização da sociedade e desenvolvimento de ferramentas de coleta e armazenamento de dados, ocorreu um aumento no volume de dados gerados (HAN *et al.*, 2011). Corrêa e Sferra (2003) descrevem a necessidade de desenvolver novas técnicas e ferramentas que sejam capazes de transformar esse grande volume de dados em informações significativas e em conhecimento. Tal conhecimento, implícito e difícil de ser identificado utilizando-se sistemas convencionais de gerenciamento de banco de dados, pode auxiliar no planejamento, gestão e tomadas de decisão. Para isso, esse trabalho utiliza técnicas de Extração de Conhecimento, e de Sistema de Recomendação para auxiliar na tomada de decisão.

A Seção 2.1 apresenta as definições de Extração de Conhecimento e as suas etapas. Posteriormente, a Seção 2.2 descreve os Sistemas de Recomendação.

## 2.1 EXTRAÇÃO DE CONHECIMENTO

Fayyad *et al.* (1996) definem a Extração de Conhecimento, do inglês *Knowledge-Discovery in Databases* ou KDD, como um processo não trivial de identificação de padrões válidos, novos, potencialmente úteis e, em última análise, compreensíveis nos dados. Kurgan e Musilek (2006) defendem que um modelo definido de processo ajuda a entender a Extração de Conhecimento e provê um guia a ser seguido nas etapas de planejamento e execução de projetos de descoberta, ilustrado na Figura 1.

**Figura 1 – Etapas do Processo de Extração de Conhecimento**




**Fonte: Romao (2016) adaptado de Fayyad *et al.* (1996).**

Fayyad *et al.* (1996) descrevem as nove etapas do processo de Extração de Conhecimento:

1. Compreender os dados já conhecidos de maneira a identificar o objetivo do processo;
2. Selecionar as variáveis ou amostras de dados em que a descoberta deverá ser realizada;
3. Pré-processar os dados, incluindo tarefas de remoção de ruídos, tratamento de dados faltantes, entre outras;
4. Reduzir os dados, realizando as transformações e consolidações necessárias, sem sacrificar a integridade dos dados;
5. Selecionar a técnica de Mineração de Dados de acordo com o interesse da descoberta;
6. Selecionar o algoritmo compatível com a técnica de Mineração de Dados escolhida, definindo os parâmetros e modelos a serem aplicados;
7. Minerar os dados aplicando métodos inteligentes para extrair conhecimento;
8. Interpretar os padrões extraídos, podendo envolver sua visualização ou até mesmo a visualização dos dados inseridos inicialmente e sua avaliação; e
9. Apresentar o conhecimento descoberto, incluindo a verificação e resolução de conflitos potenciais ou uso do conhecimento em outro sistema.

Na sequência, apresenta-se cada etapa do processo de Extração de Conhecimento detalhando a teoria das técnicas utilizadas neste projeto.

### 2.1.1 Compreensão e Seleção

Os dados podem apresentar formas e estruturas versáteis e significados bastante diferentes, como por exemplo: dados relacionados ao tempo ou de sequência; fluxos de dados; dados espaciais; dados multimídia; dados de rede; dados de transação; dados relacionais; entre outros. Deste modo, é necessário avaliar a melhor maneira de manipulá-los e extrair conhecimento destes, apesar de suas diferenças. Em um banco de dados relacional, se encontram estruturas denominadas tabelas que apresentam atributos e registros referentes a estes atributos. Os atributos ou variáveis possuem características de acordo com o seu tipo (HAN *et al.*, 2011).

Chakrabarti *et al.* (2008) apontam os tipos das variáveis mais comuns em bancos de dados:

- Nominal: Também chamada de variável categórica, apresenta a ausência de ordem, escala ou distância entre os valores que descreve.
- Ordinal: Variável categórica que possui a noção de ordem incorporada a elas.
- Numérico: Indica o valor real, ou contínuo, apresentando escala, distância e ordem entre os valores.

Keim (2002) defende que é possível utilizar técnicas de visualização dos dados para facilitar a análise inicial das variáveis. A exploração visual dos dados é especificamente útil quando se tem pouco conhecimento sobre os dados e quando a finalidade do estudo ainda não está totalmente definida. Tal prática permite que o usuário compreenda melhor os dados, propondo novas hipóteses, propiciando a etapa de pré-processamento dos dados.

### 2.1.2 Pré-processamento

Usualmente, os dados contidos em bancos de dados podem apresentar inconsistência, ruídos ou estarem incompletos, demandando a tarefa de pré-processamento a fim de torná-los adequados para serem minerados. Tais irregularidades podem ocorrer por diversos motivos como, por exemplo, falha em equipamentos, inserção duplicada de dados, estrutura de dados desenvolvidas indevidamente, exclusividade de dados agregados, valores aleatórios, entre outros. O pré-processamento abrange etapas como limpeza e preenchimento de dados faltantes.

A limpeza dos dados consiste na suavização de ruídos e remoção de anomalias. A suavização de ruídos e remoção de anomalias podem ser realizadas utilizando algoritmos complexos como Alisamento, que analisa os valores vizinhos para regular as divergências, e Regressão, que utiliza regressão linear para ajustar os valores (CHAKRABARTI *et al.*, 2008). Há também outras maneiras mais triviais de identificar os registros anômalos das variáveis, como por exemplo, verificando a frequência dos dados.

Além disso, recursos gráficos como Diagrama de Caixa e Histogramas podem ser usados para representar e comparar visualmente os grupos de dados (WILLIAMSON *et al.*, 1989). O Diagrama de Caixa utiliza a mediana, os quartis aproximados e os valores mínimos e máximos para ilustrar graficamente a dispersão e a simetria da distribuição dos valores. Já o

Histograma representa a distribuição de frequência ou a densidade das probabilidades dos dados da base. Tais visualizações permitem inferir sobre os valores discrepantes, melhorando a análise das informações quantitativas.

Dados faltantes podem ser tratados de diversas maneiras como, por exemplo: preenchendo manualmente as lacunas de dados; usando uma constante global para preencher automaticamente os valores; preenchendo as lacunas com a média; ou até mesmo utilizando o valor mais provável obtido por meio da aplicação de algoritmos de decisão (CHAKRABARTI *et al.*, 2008). Também é possível aplicar técnicas mais simples como a *listwise deletion*, em que são deletados os registros que contêm dados faltantes em alguma das variáveis dentro do modelo de interesse. Tal técnica pode ser utilizada para todas as classes de análises estatísticas, além de não requerer muito esforço computacional (ALLISON, 2001).

### 2.1.3 Transformação de Dados

Após a etapa de pré-processamento dos dados, é possível realizar a integração, generalização e modificação dos dados a fim de prepará-los para a etapa de Mineração de Dados. Nisbet *et al.* (2009) definem a integração como a tarefa de combinar dados de diferentes fontes, atentando-se à identificação de entidade para que as tuplas apresentem o esquema e correspondência corretos. Faz parte da etapa de transformação a generalização dos dados, que declara valor das variáveis para uma hierarquia de nível mais alto e a construção de atributos, em que novos atributos são criados a partir dos atributos já existentes. Também é possível citar como parte da etapa de transformação a seleção de atributos, a normalização das variáveis numéricas e a codificação dos atributos categóricos.

A seleção dos atributos é responsável por reduzir a dimensionalidade dos dados, diminuindo a complexidade dos algoritmos de Mineração de Dados, facilitando a análise dos resultados e atenuando o uso de recursos redundantes ou irrelevantes (LIU; MOTODA, 2012). As principais abordagens da seleção de atributos são a abordagem filtro, que considera somente as características dos dados, e a abordagem *wrapper*, que aplica algoritmo de mineração e utiliza seu desempenho como critério de avaliação (BORGES; NIEVOLA, 2006). Dentre os métodos de seleção de atributos com abordagem filtro, pode-se citar a técnica de classificação que utiliza a Variação de Limiar. Tal técnica utiliza como critério de eliminação dos atributos o limiar de variação, filtrando os menos relevantes (CHANDRASHEKAR; SAHIN, 2014).

Em relação à modificação dos dados, a fim de torná-los compatíveis com o algoritmo a

ser utilizado, é possível citar o tratamento de variáveis por modelos estatísticos, e a normalização de dados numéricos. Variáveis categóricas também requerem tratamento, realizando a sua codificação para dados numéricos ou variáveis indicadoras (NISBET *et al.*, 2009). A normalização dos dados, que atribui escala aos atributos, pode ser realizada aplicando a técnica de normalização Min-Max. Tal técnica realiza um processo de transformação linear dos dados, encaixando-os em uma escala de valores mínimo e máximo (SHALABI *et al.*, 2006). A equação geral que rege o funcionamento da normalização Min-Max é dada na Equação 1

$$x_{normalizado} = \frac{x - min(x)}{max(x) - min(x)} \cdot (Valor_{min} - Valor_{max}) + Valor_{min} \quad (1)$$

onde $Valor_{min}$ e $Valor_{max}$ são os valores de mínimo e máximo e $x$ é a variável que se quer normalizar.

Já o objetivo da codificação de dados categóricos é representar tais dados, que não apresentam uma ordem, como valores numéricos, buscando manter as representações no mínimo de dimensões possível (MCGINNIS *et al.*, 2018). A técnica *One-hot encoding* é a estratégia mais utilizada devido à sua simplicidade, transformando variáveis categóricas em vetores. Considerando uma variável categórica qualquer $x$, com $n$ valores distintos $x_1, x_2, ...x_n$, o *One-hot encoding* determina que cada valor de $x_i$ é um vetor $v$ onde cada componente de $v$ é zero, com exceção do i-ésimo componente (HANCOCK; KHOSHGOFTAAR, 2020).

### 2.1.4 Mineração de Dados

Goldschmidt *et al.* (2015) definem a Mineração de Dados como a principal etapa do processo de Extração de Conhecimento em que ocorre a busca efetiva por conhecimentos novos e úteis a partir dos dados. Pode-se classificar a Mineração de Dados pela sua abordagem, sendo ela de aprendizado supervisionado ou não supervisionado. No aprendizado supervisionado encontram-se atributos preditores que determinam o resultado esperado do algoritmo. No entanto, no aprendizado não supervisionado não há atributos preditores e o algoritmo identifica regularidades entre os dados a fim de agrupá-los em função de sua similaridade.

Han *et al.* (2011) descrevem as diversas técnicas da Mineração de Dados, sendo categorizadas de acordo com o objetivo da descoberta, sendo elas:

- Caracterização e Discriminação: Caracterização refere-se a sumarização das características e recursos gerais da classe de dados visada. Já a discriminação realiza comparação dos recursos das classes de dados alvo com outras classes em estudo. Para ambos os casos

utiliza-se o mesmo método, apresentando o mesmo formato de resultados (HAN *et al.*, 2011).

- Associação: Tarefa que consiste em definir a relação entre dados. Neste contexto os dados são apresentados como transações e busca-se localizar quais transações ocorrem simultaneamente dentro do conjunto de dados (ZAKI, 2000).

- Classificação e Regressão: Tarefas cujos dados são analisados e generalizados. A Classificação mapeia um conjunto de registros em um conjunto de classes. Já a Regressão mapeia os registros em um intervalo de valores reais (MICHIE *et al.*, 1994). É possível citar algoritmos de classificação como, por exemplo, Máquina de Vetores de Suporte (PLATT *et al.*, 1999), utilizado para análise de regressão; Aprendizagem por Árvores de Decisão (KOHAVI; QUINLAN, 2002), cujo objetivo é criar um modelo que preveja o valor de uma variável destino com base nas variáveis de entrada; *Perceptron Multilayers* (RUSSELL; NORVIG, 2004), algoritmo de aprendizado de classificadores binários; e *Naive Bayes* (CARUANA; NICULESCU-MIZIL, 2006), que desconsidera completamente a correlação entre as variáveis.

- Agrupamento: Caracterizada por aprendizado não-supervisionado, cuja funcionalidade é identificar grupos que descrevam os dados. Como resultado os elementos de cada grupo compartilham propriedades comuns que os distinguem de elementos nos demais grupos (HRUSCHKA; EBECKEN, 2003).

- Análise de anomalias: Tarefa que consiste em identificar registros do conjunto de dados cujas características divergem dos padrões gerais. Apresenta o resultados das análise como um valor real ou como rótulo binário para identificar as anomalias (HAN *et al.*, 2011).

Han *et al.* (2011) classificam as técnicas de Mineração de Dados em duas categorias, de acordo com a tarefa que realizam, sendo elas descritiva e preditiva. Técnicas descritivas são utilizadas quando busca-se padrões para entender melhor os dados disponíveis. Por outro lado, as técnicas preditivas auxiliam na predição de informações ainda desconhecidas. As técnicas de Regressão, Classificação, Agrupamento e Análise de Anomalias classificam-se como preditivas, enquanto que as técnicas de Caracterização, Discriminação e Associação classificam-se como descritivas.

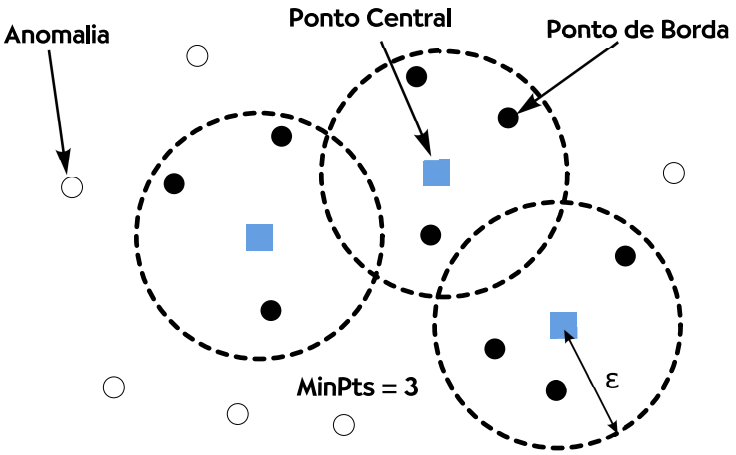
#### 2.1.4.1 Agrupamento

O método de Agrupamento, conforme definem Goldschmidt *et al.* (2015), difere do método de Classificação pelos registros não apresentarem rótulos predefinidos, sendo assim, apresentam abordagem não supervisionada. Pode fazer parte da técnica de Agrupamento determinar como será medida a similaridade entre os grupos, como codificar variáveis categóricas, como padronizar ou normalizar variáveis numéricas ou até mesmo quantos grupos serão obtidos após a mineração. Tais características dependem dos algoritmos utilizados para a aplicação do método. Os métodos de Agrupamento podem ser classificados, de acordo com o seu funcionamento, como baseados em: distância, densidade e distribuições de probabilidades.

Agrupamentos baseados em distância buscam minimizar a distância entre cada registro com a média dos grupos determinados. Entre os algoritmos baseados em distância estão: *K-Means*, *Fuzzy K-Means*, *K-Modes* e *K-Medoid* (VERMA *et al.*, 2012). Já os Agrupamentos baseados em distribuições de probabilidades consideram que todos os registros apresentam uma probabilidade diferente de zero de pertencerem a todos os agrupamentos. É possível citar, como algoritmo de Agrupamento baseado em distribuições de probabilidades, Expectativa-Maximização que apresenta duas etapas iterativas a fim de satisfazer os critérios de convergência estabelecidos. A etapa de Expectativa calcula a verossimilhança logarítmica do conjunto de dados e a etapa de Maximização calcula os parâmetros de distribuição, maximizando as probabilidade das distribuições (AGGARWAL, 2015). Por fim, os Agrupamentos baseados em densidade consideram a recorrência de registros em regiões do espaço, examinando a vizinhança da concentração e considerando-a como um grupo (GOLDSCHMIDT *et al.*, 2015). Han *et al.* (2011) citam *Density-Based Clustering Based on Connected Regions with High Density* (DBSCAN), *Ordering Points to Identify the Clustering Structure* (OPTICS) e *Clustering Based on Density Distribution Functions* (DENCLUE) como algoritmos baseados em densidade.

O DBSCAN, ilustrado na Figura 2, é um algoritmo que define os grupos, ou *clusters*, utilizando a noção de densidade definida pelo número de pontos dentro de um raio $\epsilon$ especificado, sem assumir que os grupos têm uma forma esférica. Sendo assim, cada instância é rotulada como *ponto central* se há ao menos o número mínimo, *MinPts*, de vizinhos e está dentro do raio $\epsilon$. Enquanto que as instâncias que apresentam menos vizinhos que *MinPts*, porém está dentro do raio $\epsilon$ são rotuladas Pontos de Borda. Por fim as instâncias que não seguem nenhum dos critérios são consideradas anomalias (RASCHKA, 2015).

**Figura 2 – Agrupamento com DBSCAN**




**Fonte: Adaptado de Raschka (2015).**

O algoritmo DBSCAN requer a definição dos parâmetros de distância $\epsilon$ e a quantidade mínima por grupo *MinPts*. Sander *et al.* (1998) defendem a utilização de $2 \cdot Dim$, como o melhor valor para *MinPts*, quando a base apresenta mais dimensões. Já Rahmah e Sitanggang (2016) citam a técnica K-ésimo Vizinho mais Próximo para estipular o melhor valor a ser utilizado para $\epsilon$.

Já o algoritmo de agrupamento *K-means*, método de Agrupamento baseado em distância, agrupa os registros, em *k* partes, de uma base de acordo com a média da distância entre os membros do grupo. O *K-means* define de maneira aleatória *k* centroides, sendo assim o ponto central dos grupos. Posteriormente, o algoritmo calcula as distâncias entre os registros e agrupa os dados, atribuindo-os para o centroide de menor distância. Por fim, os centroides são recalculados utilizando a média dos valores dos pontos de cada grupo (MACQUEEN *et al.*, 1967). O algoritmo requer a definição de *k*, que quando não é conhecido, pode ser obtido utilizando o Método Cotovelo, em que encontra-se o mínimo valor de *k* que representa o maior ganho no modelo.

### 2.1.5 Interpretação, avaliação e uso do conhecimento descoberto

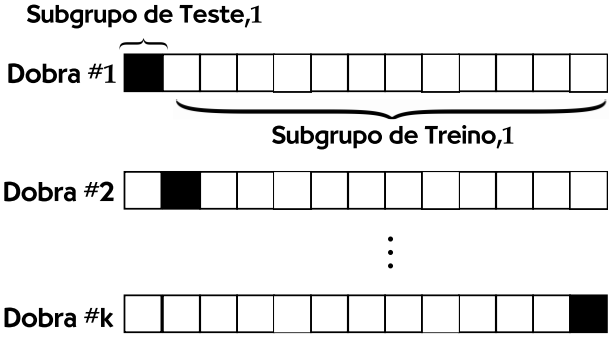
A etapa de interpretação dos padrões extraídos é realizada para analisar se o resultado é possível, internamente consistente e plausível. Se o resultado for: possível, apresentará características que condizem com a realidade; internamente consistente, não conterá informações adversárias entre si; e plausível, retrata associações coerentes (CHUNG; GRAY, 1999).

Rousseeuw (1987) aponta o uso do Coeficiente de Silhueta para interpretar e avaliar o conhecimento extraído da etapa de agrupamento utilizando o algoritmo *DBSCAN*. O método

mede a distância de cada ponto dentro de um grupo, rotulado pelo algoritmo *DBSCAN*, em comparação com as instâncias de outros grupos. O valor do Coeficiente de Silhueta varia de $-1$ a $1$, sendo que quanto mais alto o indicador, mais apropriados estão os grupos rotulados.

Já para a interpretação e validação de algoritmos de classificação, é possível utilizar a técnica de validação cruzada, ilustrada na Figura 3. Tal técnica envolve amostragem aleatória estratificada. Sendo assim, as amostras são feitas de forma que as proporções de classe nos subconjuntos individuais reflitam as proporções no conjunto de treino (HASTIE *et al.*, 2009). Na técnica de validação cruzada *K-fold* o conjunto é dividido igualmente em $k$ partes, ou dobras, atribuindo uma dessas partes como conjunto de teste e as outras $k - 1$ partes como conjunto de treinamento. O procedimento é repetido $k$ vezes, intercalando o subconjunto de treino e de teste de forma que cada um dos $k$ subconjuntos tenha servido como conjunto de validação. A precisão da técnica é dada pela média de todas as $k$ precisões obtidas nos conjuntos de validação (BERRAR, 2019).

**Figura 3 – Validação Cruzada *K-fold***




**Fonte: Adaptado de Berrar (2019).**

## 2.2 SISTEMA DE RECOMENDAÇÃO

É comum que, no dia a dia, as pessoas realizem escolhas utilizando recomendações de outras pessoas. As recomendações podem ser por meio de convívio social, avaliações de estabelecimentos ou produtos, propagandas em mídias de comunicação, entre outros. Sistemas de Recomendação podem ampliar e otimizar esse procedimento ao processar as recomendações informadas para redirecionar ao público que busca tais informações (RESNICK; VARIAN, 1997).

Adomavicius e Tuzhilin (2005) formalizam o problema de Sistemas de Recomendação definindo $C$ como um conjunto de todos os usuários e $S$ todas as possibilidades de recomendação.

Atribuiu-se $u$ como a função que mede a utilidade do item $s$ ao usuário $c$. Sendo assim, $u$ : $C \times S \rightarrow R$, sendo $R$ um conjunto ordenado. Busca-se encontrar para cada usuário $c \in C$ um item $s' \in S$ que maximiza a função de utilidade, formalizando-se na Equação 2.

$$\forall c \in C, s'_c = argmax_{s \in S} u(c,s). \tag{2}$$

Balabanović e Shoham (1997) classificam os Sistemas de Recomendação, de acordo com o tipo dos dados que analisa, como: Sistema de Recomendação Baseado em Conteúdo; Sistema de Recomendação Colaborativo; e Sistema de Recomendação Híbrido.

Os Sistemas de Recomendação Baseados em Conteúdo indicam ao usuário um item similar ao que ele já aprovou anteriormente, como demonstram Pazzani e Billsus (1997). Já Sistemas de Recomendação Colaborativos sugerem ao usuário itens que foram utilizados por outros usuários de perfil similares, conforme apresentam Goldberg *et al.* (1992). Por outro lado, Sistemas de Recomendação Híbridos utilizam a combinação dos dois métodos anteriores, atenuando as fraquezas e aproveitando as vantagens de ambos (FALK, 2019).

O objetivo da abordagem colaborativa é encontrar correspondência com base na preferência dos usuários, ponderando os interesses de usuários com perfis semelhantes para produzir uma recomendação. (TERVEEN; HILL, 2001). Cazella *et al.* (2010) definem os três passos executados por um Sistema de Recomendação Colaborativo:

- Calcular a métrica de similaridade utilizando um coeficiente de similaridade, como por exemplo, Cosseno (SINGHAL *et al.*, 2001), Euclidiana (DANIELSSON, 1980), *Pearson* (BENESTY *et al.*, 2009);
- Selecionar um subconjunto de usuários que apresentam maiores similaridades; e
- Normalizar as avaliações e computar as predições ponderando as avaliações dos vizinhos com seus pesos.

Por fim, Ricci *et al.* (2011) defendem ser necessário avaliar a qualidade e a utilidade dos Sistemas de Recomendação. Tal tarefa pode ser realizada durante variadas etapas de existência do sistema, como por exemplo durante o seu desenvolvimento em que é possível verificar a correta abordagem das recomendações. Pode-se, também, verificar a qualidade do sistema analisando a cobertura dos itens disponíveis que estão aptos a serem recomendados ao usuário. Beel *et al.* (2013) citam, no quesito utilidade de Sistemas de Recomendação, a relevância de aferir a satisfação do usuário de acordo com a recomendação obtida. A insatisfação do usuário pode

ocorrer já que nem sempre a melhor recomendação será a escolhida por ele. Sendo assim, aumentar as recomendações, por exemplo, pode melhorar a satisfação do usuário.

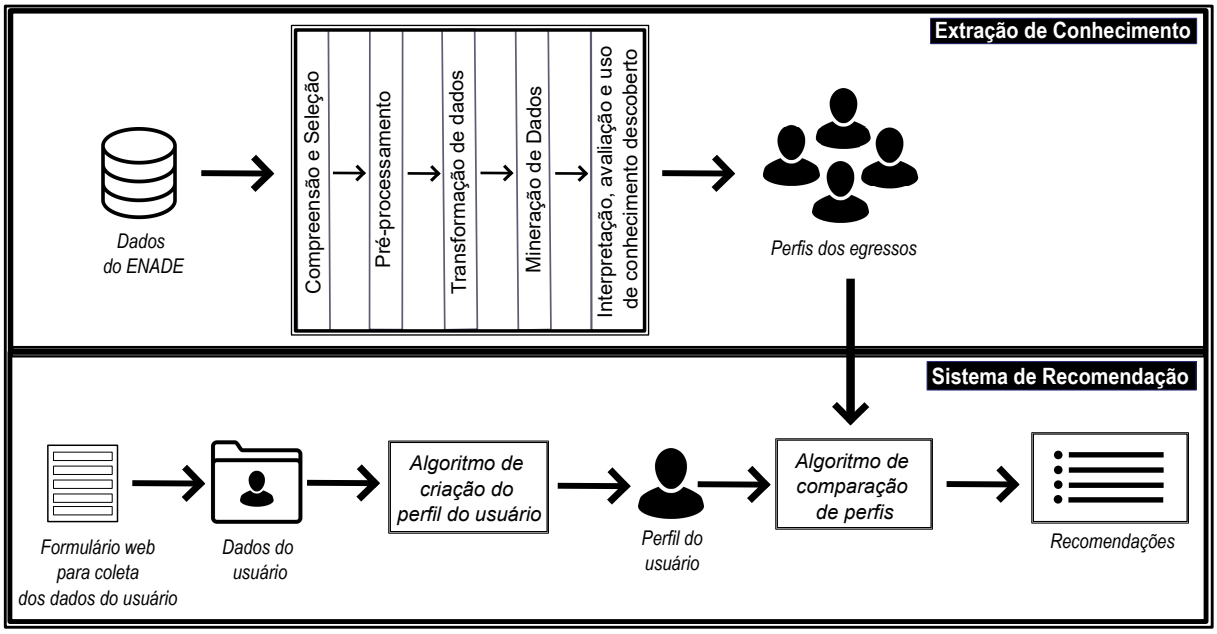
Tendo descrito os conceitos e princípios da Extração de Conhecimento e de Sistemas de Recomendação, apresenta-se o Capítulo 3 da metodologia, que descreve os métodos aplicados na execução do projeto.

# 3 MÉTODOS APLICADOS

O método de desenvolvimento deste trabalho consiste em dois processos distintos. O primeiro processo, Extração de Conhecimento, abrange as etapas necessárias para a: compreensão; seleção; pré-processamento; transformação; mineração; interpretação; e avaliação dos dados do ENADE a fim de gerar os perfis dos egressos. O processamento foi feito utilizando a linguagem de programação *Python*, por meio da aplicação *Web Jupyter Notebook*. Posteriormente, os perfis dos egressos criados foram modelados para serem utilizados no Sistema de Recomendação. Detalhes deste processo encontram-se na Seção 3.1 e apresenta-se esquematizado na parte superior da Figura 4.

O segundo processo, Sistema de Recomendações, inclui o desenvolvimento de uma aplicação *Web* responsável por coletar os dados do usuário e recomendar as instituições de ensino. O Sistema de Recomendações foi desenvolvido utilizando a linguagem de *JavaScript*, Linguagem de Marcação de Hipertexto (HTML), *Hypertext Preprocessor* (PHP), *Bootstrap* e *Cascading Style Sheets* (CSS). A aplicação permite a coleta dos dados por meio de um formulário e um algoritmo cria o perfil do usuário. O algoritmo recebe o perfil do usuário criado e realiza a comparação com os perfis de egressos a fim de fazer a recomendação. Detalhes deste processo encontram-se na Seção 3.2 e apresenta-se esquematizado na parte inferior da Figura 4.

**Figura 4 – Arquitetura proposta**




**Fonte: Autoria própria.**

Este capítulo apresenta o método utilizado para o desenvolvimento deste projeto, indicando os resultados parciais das etapas antecedentes à Mineração de Dados. A Seção 3.1 apresenta o desenvolvimento da Extração de Conhecimento e suas etapas. Já a Seção 3.2 descreve a arquitetura do Sistema de Recomendação.

## 3.1 EXTRAÇÃO DE CONHECIMENTO

### 3.1.1 Compreensão e Seleção

Os dados utilizados na pesquisa foram extraídos da base do INEP de Dados Abertos, coletados por pesquisas, avaliações e exames realizados pelo instituto (INEP, 2020b). Optou-se por utilizar os Microdados do ENADE visto que a pesquisa tem como objetivo a recomendação de instituição de ensino para ingressantes de acordo com o perfil dos egressos. No portal do INEP estão disponíveis os Microdados do ENADE separados por ano, de 2004 até 2019. Para coletar as bases do portal do INEP é necessário acessar o portal on-line, selecionar o ano que deseja detalhar e realizar o *download* do arquivo compactado com todos os arquivos necessários para uso e entendimento das bases.

Os Microdados disponibilizados acompanham um Dicionário de Variáveis que descrevem os dados, facilitando a compreensão desses. A relação detalhada das variáveis da base de cada ano podem ser encontradas no portal oficial do INEP[1] em formato de planilhas. A base dispõe de 137 variáveis sendo elas:

- Ano de realização da prova do ENADE;
- 9 variáveis com Informações da Instituição de Ensino Superior e do Curso;
- 7 variáveis com Informações do Estudante;
- 43 variáveis com Avaliação - Formação Geral e Componente Específico;
- 9 variáveis com Notas e Presença; e
- 68 variáveis de Questionário do Estudante.

Dado que a aplicação das avaliações do ENADE é feita de maneira escalonada, utilizar a base de somente um ano não abrange todos os cursos cadastrados no Ministério da Educação

[1] Disponível em: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enade

(MEC). Sendo assim, foi necessário utilizar as bases dos anos de 2017, 2018 e 2019, completando o ciclo com todos os cursos registrados e avaliados.

As bases de dados adquiridas apresentam a disposição de registros e atributos listados na Tabela 1, contendo uma diferença de atributos do ano de 2017 em relação aos outros. Tal diferença, acréscimo de 13 questões, acontece devido à adição, feita pelos organizadores do ENADE, de perguntas específicas em relação aos cursos avaliados em 2017. Realizou-se a eliminação de tais atributos a fim de igualar as condições das bases, apresentando, assim, somente 137 atributos.

**Tabela 1 – Bases adquiridas**

| Ano | Registros | Atributos |
|---|---|---|
| 2017 | 537.436 | 150 |
| 2018 | 548.127 | 137 |
| 2019 | 433.930 | 137 |

**Fonte: Fonte: Autoria Própria.**

### 3.1.2 Pré-processamento

Antes de iniciar a análise dos dados foram excluídas as variáveis irrelevantes ao estudo e as tuplas referentes aos alunos que faltaram no dia da aplicação da prova. Sendo assim, eliminou-se os atributos da base, *NU_ANO* (Ano de realização do exame), *CO_CATEGAD* (Código da categoria administrativa da IES), *CO_ORGACAD* (Código da organização acadêmica da IES), *TP_INSCRICAO_ADM* (Forma pela qual foi realizada a inscrição), *TP_INSCRICAO* (Tipo de inscrição) e *CO_REGIÃO_CURSO* (Código da região de funcionamento do curso). Visto que não há necessidade de avaliar o desempenho do acadêmico, todas as questões na seção *Avaliação - Formação Geral e Componente Específico* foram eliminadas, bem como o *Questionário de Percepção da Prova*. As questões *QE_I27* a *QE_I68* do *Questionário do Estudante* são relacionas à percepção do curso e instituição em que o acadêmico está cursando e também foram eliminadas. As bases, então, ficaram com um total de $38$ atributos cada.

Faz parte da seção *Avaliação - Formação Geral e Componente Específico* os componentes que descrevem a presença dos participantes inscritos. Sendo assim, antes de eliminar os atributos *TP_PRES* (Tipo de presença no Enade) e *TP_PR_GER* (Tipo de presença na prova), eles foram utilizados para eliminar os participantes que não estavam presentes na prova. Primeiro foram eliminados os registros dos participantes ausentes no ENADE, pelo atributo *TP_PRES*, totalizando $532.988$ registros, representando 35,0767% de toda a base. Posteriormente, eliminou-

se os registros dos participantes ausentes na prova, de acordo com o atributo *TP_PR_GER*, totalizando 2.288 registros, representando 0,2319% de toda a base. Os valores relativos a cada ano estão apresentados na Tabela 2.

**Tabela 2 – Participantes ausentes**

| Ano | *TP_PRES* | *TP_PR_GER* |
|---|---|---|
| 2017 | 87.375 (16,2577 %) | 592 (0,1315 %) |
| 2018 | 85.886 (15,6689 %) | 384 (0,0830 %) |
| 2019 | 44.087 (10,1599 %) | 88 (0,0225 %) |

**Fonte: Autoria própria.**

Após a remoção dos registros dos participantes que faltaram ao ENADE ou à prova, as bases passaram a apresentar os números de atributo e registros ilustrados na Tabela 3.

**Tabela 3 – Bases sem participantes ausentes**

| Ano | Atributos | Registros |
|---|---|---|
| 2017 | 38 | 449.469 |
| 2018 | 38 | 461.857 |
| 2019 | 38 | 389.755 |

**Fonte: Autoria própria.**

Posterior à eliminação dos registros dos participantes ausentes, verificou-se ainda restar 7.378, 2.532 e 313 registros ausentes nas bases de 2017, 2018 e 2019, respectivamente, sendo eles descritas nas tabelas 4, 5, 6.

**Tabela 4 – Variáveis ausentes em 2017**

| Atributo | Valor absoluto | Valor percentual % |
|---|---|---|
| ANO_FIM_EM | 29 | 0,00645 |
| ANO_IN_GRAD | 29 | 0,00645 |
| CO_TURNO_GRADUACAO | 376 | 0,08365 |
| QE_I01 à QE_I26 | 7.011 | 1,55984 |

**Fonte: Autoria própria.**

**Tabela 5 – Variáveis ausentes em 2018**

| Atributo | Valor absoluto | Valor percentual% |
|---|---|---|
| QE_I01 | 2.523 | 0,54627 |
| QE_I02 à QE_I25 | 2.522 | 0,54606 |
| QE_I26 | 237.619 | 51,44861 |

**Fonte: Autoria própria.**

Verificou-se, a partir da análise dos dados faltantes dos atributos cujas proporções apresentavam anormalidade, a necessidade de manipular tais dados para encaixarem-se em uma opção válida ao invés de eliminar os registros. Para as entradas nulas do atributo *QE_I16* do ano de 2019 atribui-se o valor *'9'* referente à opção válida *'Não se aplica'* e o atributo *QE_I26*,

**Tabela 6 – Variáveis ausentes em 2019**

| Atributo | Valor absoluto | Valor percentual % |
|---|---|---|
| ANO_IN_GRAD | 313 | 0,08031 |
| CO_TURNO_GRADUACAO | 18.447 | 4,73297 |
| QE_I16 | 628 | 0,16113 |

**Fonte: Autoria própria.**

que possui $51{,}4\%$ de valores ausentes no ano de 2018, foi identificado como *'Outro Motivo'*. O atributo *CO_TURNO_GRADUACAO* apresentava $18.447$ (4,7329%) de valores ausentes, sendo estes substituídos na base pelo valor $0{,}5$, a fim de manter sua característica sem a necessidade de eliminar tais registros. Posteriormente realizou-se a limpeza dos dados ausentes restantes, eliminando um total de $10.213$ registros, ao utilizar a exclusão *listwise*, ilustrados na Tabela 7.

**Tabela 7 – Registros eliminados por *listwise***

| Ano | Valor absoluto | Valor percentual % |
|---|---|---|
| 2017 | 7.378 | 1,6414 |
| 2018 | 2.523 | 0,5462 |
| 2019 | 312 | 0,0801 |

**Fonte: Autoria própria.**

Identificou-se a presença de $679$ (0,1742%) registros na base de 2019 que estavam corrompidos, sendo estes eliminados a fim de permitir as próximas etapas de análise da base. Ao analisar as proporções de respostas válidas de cada atributo, identificou-se os atributos que não apresentam variação, e portanto, não influem substancialmente as análises. Os atributos em qual a maioria dos participantes respondeu a mesma opção foram eliminados e a relação desses estão descritas no Quadro 1.

**Quadro 1 – Variáveis eliminadas**

| **Atributo** | **Descrição** | **Registros** |
|---|---|---|
| *CO_MODALIDADE* | Código da Modalidade de Ensino | *Presencial* = 85% |
| *QE_I03* | Qual a sua nacionalidade? | *Brasileira* = 97,08% |
| *QE_I12* | Ao longo da sua trajetória acadêmica, você recebeu algum tipo de bolsa de permanência? | *Nenhuma* = 78,98% |
| *QE_I13* | Ao longo da sua trajetória acadêmica, você recebeu algum tipo de bolsa acadêmica? | *Nenhuma* = 91,56% |
| *QE_I14* | Durante o curso de graduação, você participou de programas e ou atividades curriculares no exterior? | *Nenhuma* = 94,19% |
| *QE_I15* | Seu ingresso no curso de graduação se deu por meio de políticas de ação afirmativa ou inclusão social? | *Não* = 76,33% |
| *QE_I18* | Qual modalidade de ensino médio você concluiu? | *Ensino médio tradicional* = 84,94% |

**Fonte: Autoria própria.**

Além dos registros acima citados, optou-se por não utilizar o atributo *CO_CURSO*

(Código do curso no ENADE), já que apresenta $27.744$ cursos diferentes. Sendo assim, utilizou-se o registro *CO_GRUPO* (Código da Área de enquadramento do curso no ENADE), que apresenta somente $89$ respostas válidas, e representa com menor número de variáveis as indicações a serem feitas. Nesta etapa as bases de dados apresentam as características apontadas na Tabela 8.

**Tabela 8 – Base após limpeza de dados ausentes**

| Ano | Atributos | Registros |
|---|---|---|
| 2017 | 31 | 442.091 |
| 2018 | 31 | 459.334 |
| 2019 | 31 | 388.764 |

**Fonte: Autoria própria.**

Após a limpeza dos dados ausentes verificou-se a presença de anomalias. Utilizou-se um filtro de frequência, estabelecendo que os valores com frequência de reincidência $15\%$ abaixo da média deveriam ser substituídos pelo valor médio do atributo. A relação dos valores de médias usados para reposição dos dados anômalos estão descritos na Tabela 9. O Apêndice B apresenta os Diagramas de Caixa e Histogramas dos atributos *NU_IDADE*, *ANO_FIM_EM* e *ANO_IN_GRAD* antes e depois da suavização das anomalias. Por fim, os intervalos que os atributos compreendiam antes do tratamento das anomalias em comparação com os intervalos resultantes estão descritos no Quadro 2.

**Tabela 9 – Médias dos atributos com anomalias**

| Ano | NU_IDADE | ANO_FIM_EM | ANO_IN_GRAD |
|---|---|---|---|
| 2017 | 28 | 2009 | 2013 |
| 2018 | 28 | 2010 | 2015 |
| 2019 | 25 | 2011 | 2015 |

**Fonte: Autoria própria.**

**Quadro 2 – Intervalos dos atributos após tratamento de anomalias**

| **Ano** | **Atributo** | **Intervalo Anterior** | **Intervalo Posterior** |
|---|---|---|---|
| 2017 | NU_IDADE | 16-87 | 20-44 |
| | ANO_FIM_EM | 1958-2017 | 1997-2014 |
| | ANO_IN_GRAD | 1980-2017 | 2011-2015 |
| 2018 | NU_IDADE | 4-86 | 20-45 |
| | ANO_FIM_EM | 1000-2686 | 1997-2016 |
| | ANO_IN_GRAD | 1978-2099 | 2013-2017 |
| 2019 | NU_IDADE | 11-86 | 21-34 |
| | ANO_FIM_EM | 1014-2254 | 2000-2016 |
| | ANO_IN_GRAD | 3-2092 | 2013-2019 |

**Fonte: Autoria própria.**

### 3.1.3 Transformação de Dados

Todas as perguntas do Questionário do Estudante são questões objetivas de múltipla-escolha e compreendem, em média, mais de três possibilidades de respostas. Com a finalidade de simplificar tais respostas realizou-se adaptações nas opções, agrupando as análogas ou transformando as questões de maneira a admitir somente as respostas verdadeiro ou falso. As modificações realizadas nos atributos podem ser observadas no Quadro do Apêndice A, sendo que, as porcentagens das respostas para cada atributo sofrem alteração da base inicial para a etapa de transformação dos dados devido à eliminação de registros ausentes e anomalias.

De acordo com o Dicionário de Variáveis que acompanha as bases de dados é possível verificar os tipos dos atributos, que, em sua maioria, são categóricos. Os atributos *ANO_FIM_EM*, *ANO_IN_GRAD* e *NU_IDADE* são inicialmente numéricos e podem ser manipulados como tal. Já os outros atributos da base são categóricos e após realizar a codificação desses, identifica-se um crescimento indesejado no número de atributos, totalizando em $164$ variáveis. Para diminuir a dimensionalidade dos dados, realizou-se adaptações nas questões e nas respostas, agrupando-as e/ou transformando-as em variáveis numéricas quando os atributos são ordinais.

Uma das variáveis criadas foi rótulo dos registros, que representa a recomendação a ser realizada pelo sistema e consiste na combinação das variáveis *CO_GRUPO*, *CO_IES*, *CO_MUNIC_CURSO* e *CO_UF_CURSO*. Tal combinação caracteriza cada curso de cada uma das unidades de instituição de ensino superior registradas no MEC.

Após realizar as adaptações nos atributos as bases foram concatenadas, sendo possível realizar a codificação dos atributos categóricos. Para codificar as variáveis categóricas em variáveis indicadoras utilizou-se a técnica *One-hot encoding*, implementado pela função *pandas.get_dummies* da biblioteca *Pandas*. No Quadro 3 ilustram-se os atributos categóricos e suas variáveis equivalentes quando vetorizados. Além dos atributos codificados, restou na base os atributos *CO_GRUPO, NU_IDADE, TP_SEXO, ANO_FIM_EM, ANO_IN_GRAD, CO_-TURNO_GRADUACAO, QE_I07, QE_I08, QE_I10, QE_I17, QE_I21, QE_I22* e *QE_I24* que não são categóricos.

Posterior à codificação dos atributos categóricos, a base passou a ter $95$ atributos. Realizou-se, assim, a normalização das variáveis numéricas *ANO_FIM_EM*, *ANO_IN_GRAD* e *NU_IDADE*. Por fim, foi utilizado o método de Variação de Limiar, que seleciona atributos de acordo com a importância que eles apresentam ao modelo. Tal método é implementado pela

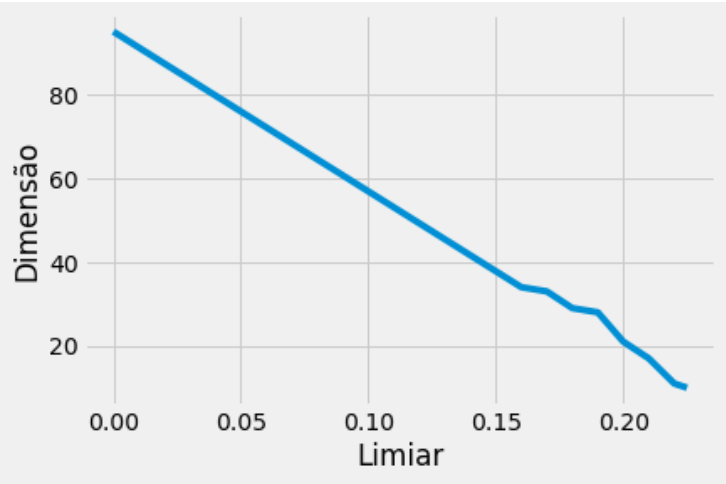
**Quadro 3 – Atributos codificados**

| Atributo Categórico | Atributo Indicador |
|---|---|
| *QE_I01* | *QE_I01_A, QE_I01_B, QE_I01_E* |
| *QE_I02* | *QE_I02_A, QE_I02_B, QE_I02_D, QE_I02_F* |
| *QE_I04* | *QE_I04_A, QE_I04_B, QE_I04_C, QE_I04_D, QE_I04_E, QE_I04_F* |
| *QE_I05* | *QE_I05_A, QE_I05_B, QE_I05_C, QE_I05_D, QE_I05_E, QE_I05_F* |
| *QE_I06* | *QE_I06_B, QE_I06_C,QE_I06_F* |
| *QE_I09* | *QE_I09_A, QE_I09_B, QE_I09_C, QE_I09_D, QE_I09_E,QE_I09_F* |
| *QE_I11* | *QE_I11_A, QE_I11_B, QE_I11_E, QE_I11_Outros* |
| *QE_I16* | *QE_I16_11.0, QE_I16_12.0, QE_I16_13.0, QE_I16_14.0 QE_I16_15.0, QE_I16_16.0, QE_I16_17.0, QE_I16_21.0 QE_I16_22.0, QE_I16_23.0, QE_I16_24.0, QE_I16_25.0 QE_I16_26.0, QE_I16_27.0, QE_I16_28.0, QE_I16_29.0 QE_I16_31.0, QE_I16_32.0, QE_I16_33.0, QE_I16_35.0, QE_I16_41.0, QE_I16_42.0, QE_I16_43.0, QE_I16_50.0, QE_I16_51.0, QE_I16_52.0, QE_I16_53.0, QE_I16_99.0* |
| *QE_I19* | *QE_I19_A, QE_I19_B, QE_I19_C* |
| *QE_I20* | *QE_I20_A, QE_I20_B, QE_I20_C, QE_I20_K* |
| *QE_I23* | *QE_I23_A, QE_I23_B, QE_I23_C, QE_I23_D, QE_I23_E* |
| *QE_I25* | *QE_I25_A, QE_I25_C, QE_I25_E, QE_I25_H* |
| *QE_I26* | *QE_I26_A, QE_I26_B, QE_I26_C,QE_I26_F, QE_I26_I* |

**Fonte: Autoria própria.**

função *sklearn.feature_selection.VarianceThreshold*, da biblioteca *Sklearn*. A função requer a definição do parâmetro de limiar, estabelecendo o limite de variação dos atributos. Foi aplicado o algoritmo alterando seu requisito a fim de controlar a sensibilidade da seleção, resultando no aumento ou diminuição da quantidade de atributos selecionados pelo algoritmo. O Quadro 4 apresenta as variáveis selecionadas de acordo com os diferentes valores de limiar inseridos na função. A Figura 5 ilustra a dimensionalidade da amostra resultante da aplicação do algoritmo em relação ao limiar utilizado.

**Figura 5 – Relação entre Limiar e Dimensionalidade da amostra resultante**




**Fonte: Autoria própria.**

Os resultados da função seletora com os diferentes valores de limiar foram armazenados

**Quadro 4 – Variáveis selecionadas pelo algoritmo de Variação de Limiar**

| Limiar | Atributos | Quantidade |
|---|---|---|
| 0,16 | *CO_GRUPO, TP_SEXO, CO_TURNO_GRADUACAO, QE_I10, QE_I17, QE_I21, QE_I24, QE_I01_A, QE_I01_B, QE_I02_A, QE_I02_D, QE_I04_B, QE_I04_D, QE_I05_B, QE_I05_D, QE_I06_B, QE_I06_C, QE_I09_B, QE_I09_C, QE_I11_A, QE_I11_B, QE_I11_E, QE_I16_35.0, QE_I19_B, QE_I20_A, QE_I20_C, QE_I20_K, QE_I23_B, QE_I23_C, QE_I25_A, QE_I25_E, QE_I25_H, QE_I26_F, QE_I26_I* | 34 |
| 0,17 | *CO_GRUPO, TP_SEXO, CO_TURNO_GRADUACAO, QE_I10, QE_I17, QE_I21, QE_I24, QE_I01_A, QE_I02_A, QE_I02_D, QE_I04_B, QE_I04_D, QE_I05_B, QE_I05_D, QE_I06_B, QE_I06_C, QE_I09_B, QE_I09_C, QE_I11_A, QE_I11_B, QE_I11_E, QE_I16_35.0, QE_I19_B, QE_I20_A, QE_I20_C, QE_I20_K, QE_I23_B, QE_I23_C, QE_I25_A, QE_I25_E, QE_I25_H, QE_I26_F, QE_I26_I* | 33 |
| 0,18 | *CO_GRUPO, TP_SEXO, CO_TURNO_GRADUACAO, QE_I10, QE_I17, QE_I21, QE_I24, QE_I01_A, QE_I02_A, QE_I02_D, QE_I04_B, QE_I04_D, QE_I05_D, QE_I06_B, QE_I06_C, QE_I09_B, QE_I09_C, QE_I11_B, QE_I11_E, QE_I19_B, QE_I20_A, QE_I20_C, QE_I20_K, QE_I23_B, QE_I23_C, QE_I25_A, QE_I25_E, QE_I25_H, QE_I26_I* | 29 |
| 0,19 | *CO_GRUPO, TP_SEXO, CO_TURNO_GRADUACAO, QE_I10, QE_I17, QE_I21, QE_I24, QE_I01_A, QE_I02_A, QE_I02_D, QE_I04_B, QE_I04_D, QE_I05_D, QE_I06_B, QE_I06_C, QE_I09_B, QE_I09_C, QE_I11_B, QE_I11_E, QE_I19_B, QE_I20_A, QE_I20_C, QE_I20_K, QE_I23_B, QE_I23_C, QE_I25_E, QE_I25_H, QE_I26_I* | 28 |
| 0,20 | *CO_GRUPO, TP_SEXO, CO_TURNO_GRADUACAO, QE_I17, QE_I21, QE_I24, QE_I02_A, QE_I02_D, QE_I04_D, QE_I05_D, QE_I06_B, QE_I06_C, QE_I09_B, QE_I11_B, QE_I11_E, QE_I19_B, QE_I20_C, QE_I23_B, QE_I23_C, QE_I25_E, QE_I26_I* | 21 |
| 0,205 | *CO_GRUPO, TP_SEXO, CO_TURNO_GRADUACAO, QE_I17, QE_I21, QE_I24, QE_I02_A, QE_I02_D, QE_I04_D, QE_I05_D, QE_I06_B, QE_I11_B, QE_I11_E, QE_I19_B, QE_I20_C, QE_I23_B, QE_I23_C, QE_I25_E, QE_I26_I* | 19 |
| 0,21 | *CO_GRUPO, TP_SEXO, CO_TURNO_GRADUACAO, QE_I17, QE_I21, QE_I24, QE_I02_A, QE_I02_D, QE_I04_D, QE_I05_D, QE_I06_B, QE_I11_B, QE_I19_B, QE_I20_C, QE_I23_B, QE_I25_E, QE_I26_I* | 17 |
| 0,215 | *CO_GRUPO, TP_SEXO, CO_TURNO_GRADUACAO, QE_I21, QE_I02_A, QE_I02_D, QE_I04_D, QE_I05_D, QE_I06_B, QE_I19_B, QE_I20_C, QE_I23_B, QE_I25_E, QE_I26_I* | 14 |
| 0,22 | *CO_GRUPO, TP_SEXO, QE_I02_A, QE_I02_D, QE_I05_D, QE_I06_B, QE_I19_B, QE_I20_C, QE_I23_B, QE_I25_E, QE_I26_I* | 11 |
| 0,225 | *CO_GRUPO, TP_SEXO, QE_I02_A, QE_I06_B, QE_I19_B, QE_I20_C, QE_I23_B, QE_I26_I* | 10 |

**Fonte: Autoria própria.**

a fim de realizar o processo de Mineração de Dados com tais opções e realizar a comparação em relação à acurácia da aplicação.

### 3.1.4 Mineração de Dados

Inicialmente, realizou-se a aplicação dos algoritmos utilizando uma amostra da base de dados, a fim de testar sua eficácia em variadas condições. Em primeiro momento utilizou-se um

subconjunto com amostragem aleatória de $10\%$ da base. Posteriormente, realizou-se a partição da base utilizando o atributo *CO_GRUPO* como medida. O particionamento da base pelo curso permitiu aumentar a fidelidade na representação dos dados. Além disso, auxilia na obtenção de resultados mais compatíveis com o cenário do objetivo do trabalho, em que serão feitas as análises dos perfis de acordo com o curso escolhido pelo usuário. Sendo assim, descartou-se a utilização de amostragem aleatória para a aplicação dos algoritmos. As análises dos perfis de acordo com o curso foram realizadas empregando técnicas de agrupamento e classificação, permitindo a comparação entre os resultados obtidos.

Foi utilizada a técnica de Agrupamento com a finalidade de gerar grupos de acordo com os perfis dos participantes. Para tal, utilizou-se o DBSCAN, baseado em densidade. O algoritmo DBSCAN está implementado pela função *sklearn.cluster.DBSCAN*, da biblioteca *Sklearn*. Por padrão a função tem configurado os parâmetros como: $\epsilon$ = 0.5, Pontos Mínimos = 5 e Métrica = Euclidiana. O valor de pontos mínimos utilizado foi calculado a partir de $2 \cdot Dim$, havendo assim, variação no valor escolhido de acordo com a dimensão da base resultante da seleção de variáveis. Verificou-se que o melhor limiar a escolher, ou seja, o que seleciona o melhor número de atributos a serem agrupados pelo algoritmo DBSCAN, equivale a $0{,}215$ (14 atributos). Ao calcular a distância $\epsilon$, aplicando o algoritmo de K-ésimo Vizinho mais Próximo, obteve-se um valor inviável. Sendo assim, o valor de $\epsilon$ foi escolhido empiricamente, aplicando a técnica de agrupamento e variando-o, obtendo, assim, o melhor valor possível a ser utilizado.

Realizou-se, também, a modificação do parâmetro de *MinPts*, a fim de diminuir o número de amostras rotuladas como anomalias. A Figura 6, presente na Seção 4.1, retrata os resultados das modificações realizadas, com número de grupos gerados e o número de amostras rotuladas como anomalias.

Posteriormente, aplicou-se o algoritmo *K-means*, implementado pela função *sklearn.cluster.KMeans*, da biblioteca *Sklearn*. Tal função requer a indicação do número de grupos a ser formado, sendo assim, foi definido como o número de opções de instituições contido em cada grupo. Também requer o número máximo de iterações a ser feito em cada operação, definido como *300*, e o número de vezes que o algoritmo será executado com diferentes atribuições de partida aleatória, definido como *20*. A Figura 9, presente na Seção 4.1, apresenta os resultados obtidos utilizando um curso que representa $10\%$ da base de dados, variando o limiar escolhido na etapa de seleção de variáveis. A implementação da técnica foi desenvolvida em duas partes, sendo a primeira o agrupamento, utilizando *K-means*, e a segunda a classificação da instituição,

utilizando os subconjuntos de treino e teste, resultado das partições da validação cruzada *K-Fold*. Para cada instância do subconjunto de teste identificou-se qual dos centroides formados pelo *K-means* apresentava maior similaridade com a instância. Posteriormente, verificou-se, dentro do grupo ao qual pertencia tal centroide, quais os alunos que apresentavam perfil mais similar à instância de teste.

Por fim, aplicou-se um algoritmo supervisionado de classificação baseada em similaridade de vetores para rotular as instituições de ensino, utilizando subconjuntos de teste e treino, conforme as partições da validação cruzada *K-Fold*. O processo de classificação consistiu na identificação dos perfis predominantes de cada curso e de cada instituição, contabilizando quantos perfis diferentes existiam para cada instituição e eliminando os que menos repetiam-se. Posteriormente, realizou-se a comparação de cada amostra do subconjunto de teste com os perfis predominantes anteriormente definidos, indicando o que apresentava maior similaridade. O resultado do desempenho da técnica utilizando um curso que representa $10\%$ da base estão apresentados na Figura 9, presentes na Seção 4.1. Já a média de desempenho dentro de cada subconjunto e, assim, a média do desempenho geral da base, podem ser observadas no Quadro 5, presente na Seção 4.1.

### 3.1.5 Interpretação e avaliação do conhecimento descoberto

Durante o desenvolvimento do trabalho foram utilizadas o Coeficiente de Silhueta e de validação cruzada *K-fold*. A técnica de Coeficiente de Silhueta foi aplicada a fim de validar o conhecimento extraído ao utilizar o algoritmo *DBSCAN* para obter os perfis dos egressos.

Já a técnica de validação cruzada *K-fold* foi utilizada para avaliar e interpretar o algoritmo supervisionado de classificação, além da classificação realizada posterior ao agrupamento *K-means*. Foi definida a quantidade de 10 dobras para realizar a amostragem dos subconjuntos de teste e treino. Nesta etapa a base está unificada e apresenta $1.290.189$ registros, sendo que os atributos variam de acordo com o limiar escolhido na etapa de seleção de variáveis. A precisão da técnica, medida pela média das precisões obtidas nos conjuntos de validação, pode ser observada no Quadro 5, presente na Seção 3.1.4.

## 3.2 SISTEMA DE RECOMENDAÇÃO

O Sistema de Recomendação é uma aplicação *Web*, desenvolvida utilizando *JavaScript*, HTML, PHP, *Bootstrap* e CSS, contendo um formulário on-line. Utilizou-se arquivos *Comma-Separated Values* (CSV), que podem ser facilmente acessados por *Javascript*, para armazenar os perfis resultantes do processo de Extração de Conhecimento. Sendo assim, não há a necessidade de vincular a aplicação a um gerenciador de banco de dados. A coleta dos dados é realizada por meio do formulário e um algoritmo produz a estrutura do perfil do usuário, fazendo a manipulação das informações inseridas. Dentre as questões presentes no formulário tem-se o campo de seleção do curso objetivo do acadêmico, utilizado pelo sistema para filtrar os perfis a serem comparados internamente.

Após a inserção dos dados, o Sistema de Recomendação realiza a comparação do perfil estabelecido em relação aos perfis específicos dos egressos, de acordo com o curso que foi informado pelo usuário. O algoritmo *distance*, implementado na biblioteca *scipy.spatial*, que calcula a distância entre duas amostras, de acordo com a métrica configurada. Como métrica de distância utilizou-se a Euclidiana. O algoritmo, então, é responsável por medir a distância entre o perfil do usuário e os perfis dos egressos do requerido curso. Por fim, o Sistema de Recomendação recomenda as três instituições que apresentam perfis com menor distância, ou seja, maior similaridade, em relação ao perfil do usuário. As telas de formulário e a representação da recomendação de instituições, resultante da utilização do sistema, estão ilustradas nas figuras 12 e 13, presentes na Seção 4.2.

# 4 RESULTADOS

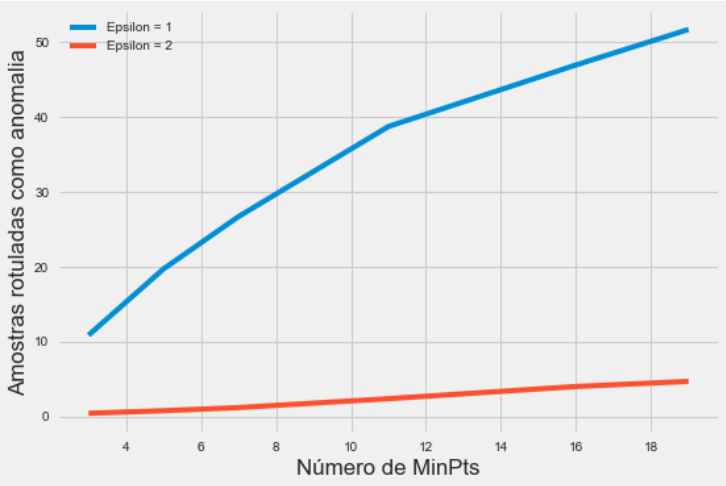
Este capítulo apresenta e discute os resultados obtidos a partir do desenvolvimento do projeto. A Seção 4.1 apresenta os resultados da aplicação das etapas de Extração de Conhecimento, com ênfase nos resultados da etapa de Mineração de Dados. Já a Seção 4.2 descreve a aplicação que compreende o Sistema de Recomendação.

## 4.1 EXTRAÇÃO DE CONHECIMENTO

As etapas de Extração de Conhecimento antecedentes à etapa de Mineração de Dados resultou em uma base sem dados ausentes, sem anomalias, com variáveis categóricas codificadas em indicadores e com valores numéricos normalizados. Sendo assim, a base de dados apresentou-se em condições propícias para a aplicação das técnicas de Mineração de Dados escolhidas. O procedimento de seleção das variáveis resultou na diminuição da dimensionalidade da base de dados, conforme apresentado na Figura 4. A partir disso, aplicou-se os algoritmos de agrupamento e classificação na base, utilizando as diferentes opções de conjunto de variáveis disponíveis.

Como é possível observar na Figura 7, escolher o valor *MinPts* utilizando a regra $2 \cdot Dim$ não gerou uma quantidade aceitável de grupos. Sendo que, quanto maior *MinPts*, menor o número de grupos e maior a proporção de amostras rotuladas como anomalias, conforme ilustram as figuras 6 e 7.

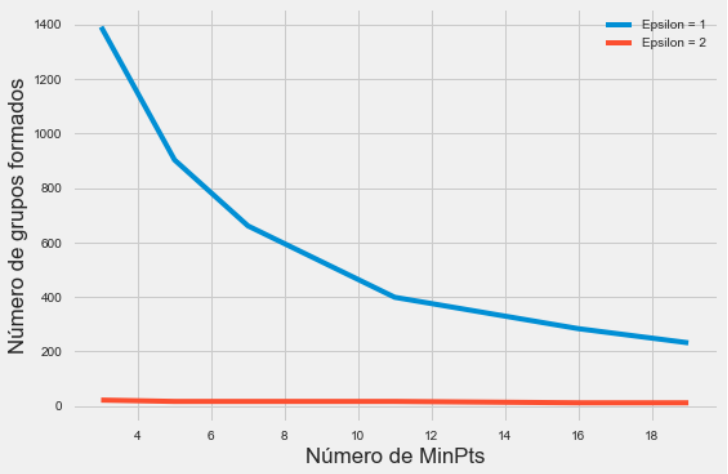
**Figura 6 – Relação entre $\epsilon$, *MinPts* e Amostras rotuladas como anômalas**




**Fonte: Autoria própria.**

Ao analisar os grupos gerados, em relação à instituição que as instâncias pertencem,

**Figura 7 – Relação entre $\epsilon$, *MinPts* e Número de grupos formado**

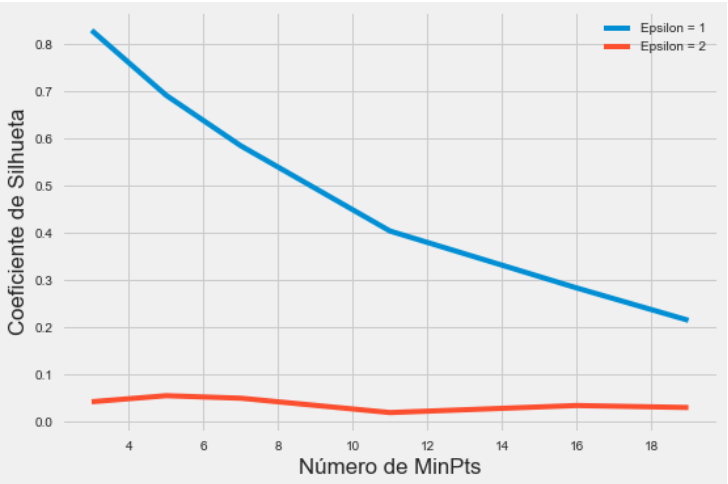



**Fonte: Autoria própria.**

verificou-se não ser possível associar cada grupo a uma instituição. Os grupos formados, por mais que numerosos, encontravam-se desbalanceados, ou seja, haviam muitas instâncias em alguns poucos grupos, enquanto que, maior parte dos grupos continham menos de 10 instâncias cada. Além disso, os grupos não apresentavam uma instituição predominante, ou seja, haviam poucas instâncias em cada grupo que possuíam a mesma instituição.

A validação do método de agrupamento foi feita ao calcular o Coeficiente de Silhueta, ilustrado na Figura 8, e ao analisar se os resultados são internamente consistentes. A partir desses dois critérios é possível inferir que os resultados não foram satisfatórios, não sendo viável utilizar o conhecimento extraído no Sistema de Recomendação.

**Figura 8 – Relação entre $\epsilon$, *MinPts* e Coeficiente de Silhueta**

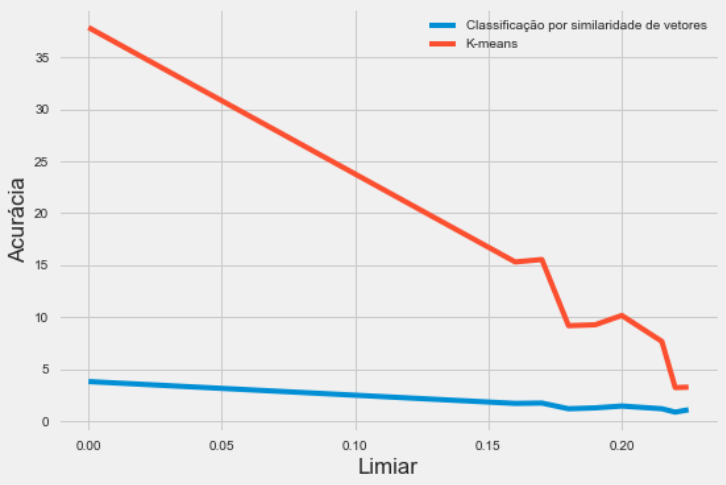



**Fonte: Autoria própria.**

Após a aplicação do algoritmo *DBSCAN*, implementou-se o algoritmo *K-means*, de

agrupamento por distância. Aplicou-se o algoritmo a um curso que representa $10\%$ da base e variou-se o limiar, que infere na dimensionalidade da base. Observou-se que, quanto menor o limiar, e consequentemente maior a dimensionalidade da base, maior a acurácia do algoritmo, conforme ilustra a Figura 9. Sendo assim, optou-se por não utilizar a seleção de variáveis para a execução da tarefa em toda a base de dados.

**Figura 9 – Relação de acurácia média dos algoritmos de classificação e *K-means* de acordo com limiar**




**Fonte: Autoria própria.**

O método de classificação por similaridade de perfis, se for analisado somente a média de acerto geral dos cursos, não apresentou resultados aceitáveis, conforme apresentado no Quadro 5, já que a taxa de acerto médio ficou próxima de $5{,}5\%$, apresentando coeficientes de similaridade baixos para as recomendações realizadas pelo sistema. Já o algoritmo que utiliza agrupamento *K-means* apresentou um melhor desempenho médio, próximo a $27{,}9\%$ de taxa de acerto. Mesmo assim, o algoritmo não apresentaria a precisão necessária para realizar recomendações com segurança.

**Quadro 5 – Resultados obtidos pelos algoritmos de classificação por similaridade de vetores e *K-means***

| Algoritmo | **Média de Acurácia (%)** | **Medida de Distância** | **Tempo de processamento** |
|---|---|---|---|
| Classificação por Similaridade | 5,5183 | 3,1355 | 7h 02min |
| *K-means* | 27,9312 | 2,8624 | 7h 34min |

**Fonte: Autoria própria.**

Porém, é necessário avaliar a taxa de acerto média de cada curso, levando em consideração o número de instâncias e o número de instituições que cada curso apresenta. Tal análise estatística permite verificar o comportamento dos algoritmos, avaliando se as taxas de acerto estão compatíveis com o cenário em que se encontra. Os quadros 6 e 7 apresentam o

desvio padrão, a média, o mínimo e o máximo das taxas de acerto, quantidade de instituições e quantidade de instâncias por curso.

**Quadro 6 – Resultados obtidos pelo algoritmo de classificação por similaridade de vetores**

| | **Taxa de Acerto por curso (%)** | **Quantidade de instituições** | **Quantidade de alunos por curso** |
|---|---|---|---|
| Desvio padrão | 3,2698 | 237,5254 | 23.546,8929 |
| Média | 5,5183 | 220,0786 | 14.496,5056 |
| Mínimo | 1,0453 | 18 | 388 |
| Máximo | 17,3035 | 1.423 | 12.6627 |

**Fonte: Autoria própria.**

**Quadro 7 – Resultados obtidos pelo algoritmo de classificação com *K-means***

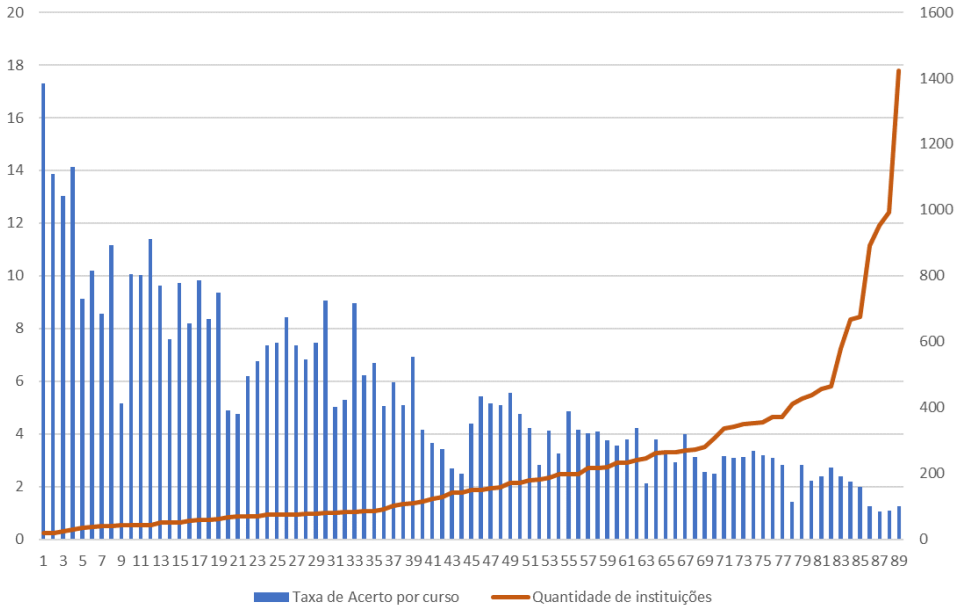
| | **Taxa de Acerto por curso (%)** | **Quantidade de instituições** | **Quantidade de alunos por curso** |
|---|---|---|---|
| Desvio padrão | 9,4441 | 198,3383 | 23.546,8939 |
| Média | 27,9312 | 177,4269 | 14.496,5056 |
| Mínimo | 8,6385 | 12 | 388 |
| Máximo | 52,9371 | 1.172 | 12.6627 |

**Fonte: Autoria própria.**

Já as figuras 10 e 11 apresentam a relação entre o número de instituições e a taxa de acerto de cada grupo. É possível verificar que a taxa de acerto diminui conforme aumenta o número de instituições (rótulos), de acordo com o esperado. Sendo assim, tem-se uma taxa de acerto aceitável, considerando a proporção da quantidade de rótulos disponíveis na base em relação ao número de instituições a serem classificadas. Tem-se, por exemplo, a maior taxa de acerto no algoritmo de classificação por similaridade de vetores (17,30%), em que o algoritmo teve que classificar um número pequeno de instituições (18). Da mesma maneira que, a menor taxa de acerto (1,25%) resultou da tentativa do algoritmo de classificar 1.423 instituições pertencentes ao curso utilizado.

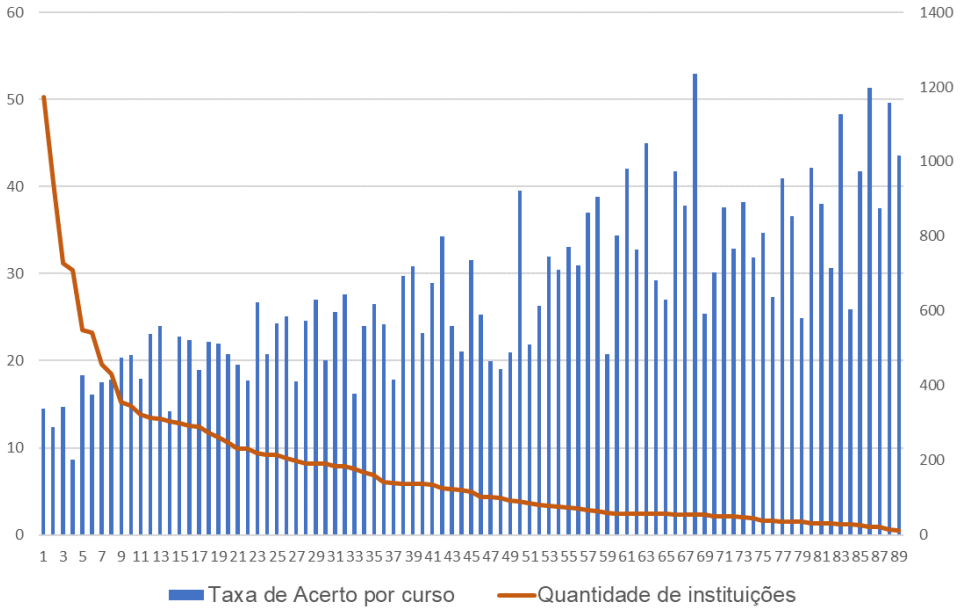
O algoritmo de classificação que utiliza o *K-means* para agrupar os perfis apresentou comportamento análogo, porém com maior desempenho. Sendo que, a menor taxa de acerto (8,63%) resultou da tentativa de classificar 709 instituições. Assim como, o curso que apresentou menor número de instituições a classificar resultou em uma taxa de acerto de 43,57%. O Anexo C apresenta as informações: média de acurácia, número de registros e número de instituições (rótulos) de cada um dos cursos.

**Figura 10 – Relação relação entre número de instituições e taxa de acerto de acordo com o algoritmo de classificação por similaridade de vetores**




**Fonte: Autoria própria.**

**Figura 11 – Relação relação entre número de instituições e taxa de acerto de acordo com o algoritmo de classificação que utiliza *K-means***




**Fonte: Autoria própria.**

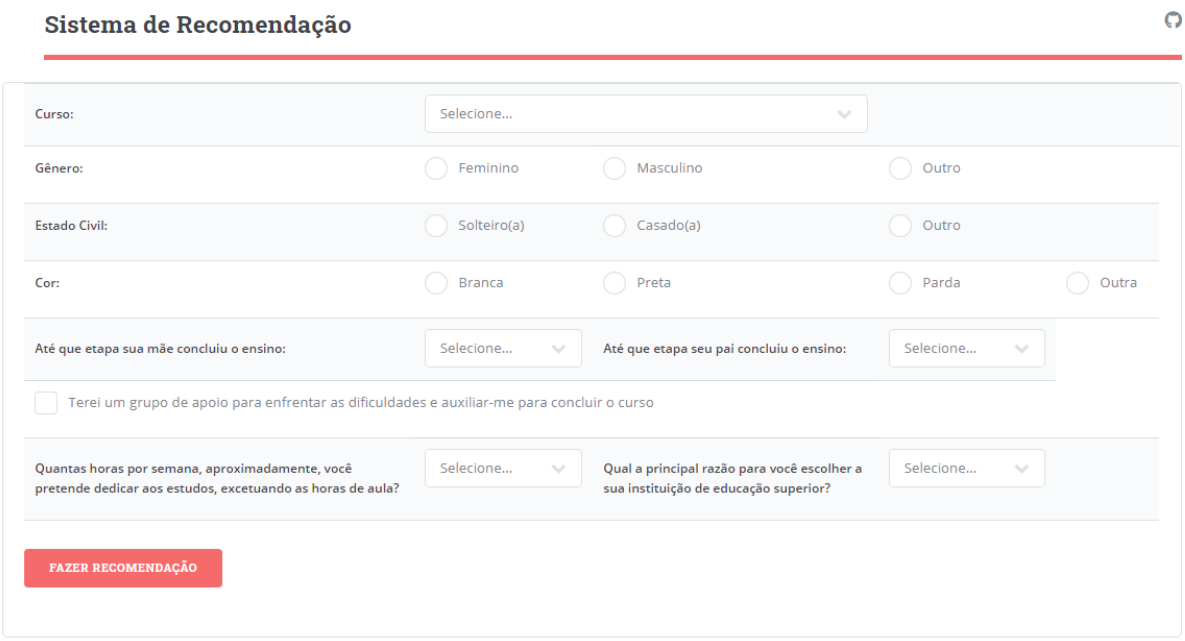
## 4.2 SISTEMA DE RECOMENDAÇÃO

A tela principal do Sistema de Recomendação possui um formulário com os campos necessários para a criação do perfil do usuário, sendo assim, os campos podem variar de acordo com os atributos selecionados. No formulário não há nenhum campo em que o usuário terá que

digitar uma informação, todas as questões estão dispostas com botões de opção, caixas de seleção e listagem. Todos os campos são obrigatórios e após o preenchimento dos dados o usuário pode solicitar a recomendação.

Após o processamento do sistema, é exibida ao usuário uma tabela com as três recomendações de instituição, contendo as informações de cidade, estado e a razão de similaridade daquela recomendação.

**Figura 12 – Tela Principal do Sistema de Recomendação**




**Fonte: Captura de tela do Sistema de Recomendação.**

**Figura 13 – Tela de resultado das recomendações**

| Instituição de Ensino Superior | Cidade | Estado | Similaridade (%) |
|---|---|---|---|
| U. FEDERAL DO RIO DE JANEIRO | RIO DE JANEIRO | Rio de Janeiro (RJ) | 28.86751345948129 |
| U. FEDERAL DE SERGIPE | SAO CRISTOVAO | Sergipe (SE) | 27.735009811261456 |
| U. FEDERAL DO AMAPÁ | MACAPA | Amapa (AP) | 28.86751345948129 |

**Fonte: captura de tela do Sistema de Recomendação.**

# 5 CONCLUSÃO

A evasão no ensino superior, desperdício social, acadêmico e econômico, têm sido tema de estudos, com o objetivo de identificar fatores que levam à evasão e como atenuá-la. Neste trabalho foi proposto o desenvolvimento de um Sistema de Recomendação de acordo com os perfis dos egressos das instituições de ensino superior. Tal sistema visa auxiliar a escolha de universidade de um futuro acadêmico, buscando assim atenuar a evasão do ensino superior. Para a análise dos perfis aplicou-se técnicas de Extração de Conhecimento utilizando as bases de dados do ENADE, fornecidas pelo INEP.

O processo de Extração de Conhecimento foi composto de cinco etapas. Na etapa de compreensão e seleção realizou-se a coleta das bases no portal do INEP e a análise dos atributos das bases conforme o Dicionário de Variáveis que as acompanha. Posteriormente, na etapa de pré-processamento, realizou-se o tratamento dos dados ausentes, anômalos e a exclusão de atributos irrelevantes. Já na etapa de transformação dos dados, os atributos foram modificados, diminuindo as opções válidas das questões e verificando os atributos que apresentam pouca variação. Nesta etapa também foi realizada a codificação das variáveis categóricas, tipo mais comum na base de dados utilizada, para variáveis indicadoras, e analisou-se a disposição dos registros de cada atributo. Após a execução da função de seleção de atributos foram evidenciados os atributos que possibilitariam um melhor resultado de mineração.

Na etapa de Mineração de Dados aplicou-se a técnica DBSCAN de agrupamento, além da técnica de classificação utilizando similaridade de vetores, para extrair os perfis dos egressos. A aplicação da técnica de agrupamento, após a diversificação dos parâmetros, resultou em grupos que rotulavam como anomalia cerca de $20\%$ da base e $0{,}8$ de Coeficiente de Silhueta. Porém, ao analisar cada um dos grupos, verificou-se que não é possível associar cada grupo a uma instituição devido à característica dos grupos de apresentarem poucas instâncias em cada grupo que possuíam a mesma instituição. Já o algoritmo de agrupamento *K-means* apresentou desempenho de $27{,}9\%$ de acurácia média, resultado melhor comparado ao algoritmo de classificação por similaridade de perfis que apresentou acurácia média de $5{,}5\%$. É possível verificar, ao analisar os resultados para cada um dos cursos, que a taxa de acerto diminui conforme aumenta o número de instituições (rótulos), de acordo com o esperado. Sendo assim, tem-se uma taxa de acerto aceitável, considerando a proporção da quantidade de rótulos disponíveis na base em relação ao número de instituições a serem classificadas.

O produto final desse trabalho foi um Sistema de Recomendação, uma ferramenta *Web* a ser utilizada pelos futuros ingressantes do ensino superior. Devido ao baixo índice de acurácia do algoritmo de classificação, o sistema apresenta recomendações de instituições que possuem baixa similaridade com o perfil do usuário. Sendo assim, o sistema pode ser utilizado como ferramenta coadjuvante na tomada de decisão, podendo ser associado a outros métodos a fim de aumentar a efetividade das recomendações.

Sugere-se, como trabalho futuro, aprimorar o Sistema de Recomendação de maneira a aumentar a sua precisão. Para tal, será necessário considerar outros métodos de análises como Sistemas Especialistas ou Aprendizagem Profunda, em que modelos prontos podem ser utilizados por meio de técnicas de Transferência por Aprendizagem. A fim de otimizar os resultados obtidos pelo algoritmo DBSCAN, sugere-se o uso de técnicas de balanceamento de dados, permitindo a formação de grupos mais distribuídos e adequados para análises. Poderá, também, ser consideradas outras fontes de dados para incorporar às bases já existentes, aumentando os atributos utilizados para a classificação das instituições. Essas bases podem ser de uso interno de instituições ou até mesmo informações oriundas de pesquisas de campo. Além disso, é possível citar melhorias que poderão ser feitas no Sistema de Recomendação, como por exemplo, apresentando ao usuário filtros e pesos a serem conferidos aos atributos, além de disponibilizar mais informações no momento da recomendação.

As informações disponibilizadas poderiam prover maior entendimento sobre a situação da instituição recomendada, como por exemplo as informações sobre a cidade em que localiza-se a instituição recomendada, a distância entre a cidade da instituição e a cidade de residência do usuário, as notas de avaliação realizadas pelo MEC em relação à instituição, uma visão resumida da opinião sobre a instituição informada pelos egressos, entre outros. Por fim, sugere-se a verificação da satisfação dos usuários em relação ao Sistema de Recomendação, apresentando-o a um grupo de futuros acadêmicos que poderá avaliar a recomendação obtida. Tal recomendação poderá ser feita, por exemplo, adicionando a possibilidade do usuário definir se concorda ou não com a recomendação. Armazenando essa avaliação dos usuários e a opção de curso escolhida, seria possível verificar para quais cursos o modelo gerado apresenta resultados mais próximos à expectativa dos usuários.

# REFERÊNCIAS


ADOMAVICIUS, Gediminas; TUZHILIN, Alexander. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. **IEEE transactions on knowledge and data engineering**, IEEE, v. 17, n. 6, p. 734–749, 2005.

AGGARWAL, Charu C. **Data mining: the textbook**. *[S.l.]*: Springer, 2015.

ALLISON, Paul D. **Missing data**. *[S.l.]*: Sage publications, 2001.

ALMEIDA, João Antonio da Silva; BRUM, Evanisa Helena Maio de; FRANCO, Márcia Elisabete Wilke; RIBEIRO, Daniela Pereira; MARTINS, Leticia Wilke Franco; MOURA, Eliane Rosa Pereira de. Gestão da vida acadêmica: Uma proposta de intervenção para diminuir a evasão no ensino superior. **RACE-Revista de Administração do Cesmac**, v. 5, p. 289–313, 2019.

BALABANOVIĆ, Marko; SHOHAM, Yoav. Fab: content-based, collaborative recommendation. **Communications of the ACM**, ACM New York, NY, USA, v. 40, n. 3, p. 66–72, 1997.

BARBARÁ, Daniel; JAJODIA, Sushil. **Applications of data mining in computer security**. *[S.l.]*: Springer Science & Business Media, 2002. v. 6.

BEEL, Joeran; LANGER, Stefan; GENZMEHR, Marcel; GIPP, Bela; BREITINGER, Corinna; NÜRNBERGER, Andreas. Research paper recommender system evaluation: a quantitative literature survey. *In*: **Proceedings of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation**. *[S.l.: s.n.]*, 2013. p. 15–22.

BENESTY, Jacob; CHEN, Jingdong; HUANG, Yiteng; COHEN, Israel. Pearson correlation coefficient. *In*: **Noise reduction in speech processing**. *[S.l.]*: Springer, 2009. p. 1–4.

BERRAR, Daniel. Cross-validation. **Encyclopedia of bioinformatics and computational biology**, Academic, v. 1, p. 542–545, 2019.

BIAZUS, Cleber Augusto *et al.* Sistema de fatores que influenciam o aluno a evadir-se dos cursos de graduação na ufsm e na ufsc: um estudo no cursos de ciências contábeis. Florianópolis, SC, 2004.

BORGES, HELYANE BRONOSKI; NIEVOLA, JC. **Redução de Dimensionalidade em Bases de Dados de Expressão Gênica**. 2006. Tese (Doutorado) — Dissertação de Mestrado, PPGIa-PUCPR, 2006.

CAMILO, Cássio Oliveira; SILVA, João Carlos da. Mineração de dados: Conceitos, tarefas, métodos e ferramentas. **Universidade Federal de Goiás (UFC)**, p. 1–29, 2009.

CARUANA, Rich; NICULESCU-MIZIL, Alexandru. An empirical comparison of supervised learning algorithms. *In*: **Proceedings of the 23rd international conference on Machine learning**. *[S.l.: s.n.]*, 2006. p. 161–168.

CAZELLA, Sílvio César; NUNES, MASN; REATEGUI, Eliseo. A ciência da opinião: Estado da arte em sistemas de recomendação. **André Ponce de Leon F. de Carvalho; Tomasz Kowaltowski..(Org.). Jornada de Atualização de Informática-JAI**, p. 161–216, 2010.

CHAKRABARTI, Soumen; COX, Earl; FRANK, Eibe; GÜTING, Ralf Hartmut; HAN, Jiawei; JIANG, Xia; KAMBER, Micheline; LIGHTSTONE, Sam S; NADEAU, Thomas P; NEAPOLITAN, Richard E *et al.* **Data mining: know it all**. *[S.l.]*: Morgan Kaufmann, 2008.

CHANDRASHEKAR, Girish; SAHIN, Ferat. A survey on feature selection methods. **Computers & Electrical Engineering**, Elsevier, v. 40, n. 1, p. 16–28, 2014.

CHUNG, H Michael; GRAY, Paul. Data mining. **Journal of management information systems**, Taylor & Francis, v. 16, n. 1, p. 11–16, 1999.

CORRÊA, Â; SFERRA, HH. Conceitos e aplicações de data mining. **Revista de ciência & tecnologia**, v. 11, p. 19–34, 2003.

Dados do Censo da Educação Superior. **divulgado pelo Ministério da Educação (MEC) e pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), em 2014**. 2016.

DANIELSSON, Per-Erik. Euclidean distance mapping. **Computer Graphics and image processing**, Elsevier, v. 14, n. 3, p. 227–248, 1980.

DAVOK, Delsi Fries; BERNARD, Rosilane Pontes. Avaliação dos índices de evasão nos cursos de graduação da universidade do estado de santa catarina-udesc. **Avaliação: Revista da Avaliação da Educação Superior (Campinas)**, SciELO Brasil, v. 21, n. 2, p. 503–522, 2016.

DURAND, G; LAPLANTE, F; KOP, R. A learning design recommendation system based on markov decision processes. *In*: **KDD 2011 Workshop: Knowledge Discovery in Educational Data, ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2011) in San Diego, CA**. *[S.l.: s.n.]*, 2011.

FALK, Kim. **Practical recommender systems**. *[S.l.]*: Manning Publications, 2019.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From data mining to knowledge discovery in databases. **AI magazine**, v. 17, n. 3, p. 37–37, 1996.

FILHO, Roberto Leal Lobo Silva; MOTEJUNAS, Paulo Roberto; HIPÓLITO, Oscar; LOBO, Maria Beatriz de Carvalho Melo. A evasão no ensino superior brasileiro. **Cadernos de pesquisa**, SciELO Brasil, v. 37, n. 132, p. 641–659, 2007.

GIUDICI, Paolo; FIGINI, Silvia. **Applied data mining for business and industry**. *[S.l.]*: Wiley Online Library, 2009.

GOLDBERG, David; NICHOLS, David; OKI, Brian M; TERRY, Douglas. Using collaborative filtering to weave an information tapestry. **Communications of the ACM**, ACM New York, NY, USA, v. 35, n. 12, p. 61–70, 1992.

GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel; BEZERRA, Eduardo. **Data Mining**. *[S.l.]*: Elsevier Brasil, 2015.

HAN, Jiawei; PEI, Jian; KAMBER, Micheline. **Data mining: concepts and techniques**. *[S.l.]*: Elsevier, 2011.

HANCOCK, John T; KHOSHGOFTAAR, Taghi M. Survey on categorical data for neural networks. **Journal of Big Data**, Springer, v. 7, p. 1–41, 2020.

HASEBROOK, Joachim P; NATHUSIUS, Wolfgang. An expert advisor for vocational guidance. **Journal of Artificial Intelligence in Education**, AACE ASSOCIATION FOR THE ADVANCEMENT OF, v. 8, p. 21–42, 1997.

HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. **The elements of statistical learning: data mining, inference, and prediction**. *[S.l.]*: Springer Science & Business Media, 2009.

HRUSCHKA, Eduardo R; EBECKEN, Nelson FF. A genetic algorithm for cluster analysis. **Intelligent Data Analysis**, IOS Press, v. 7, n. 1, p. 15–25, 2003.

INEP. **Apresentação Exame Nacional de Desempenho dos Estudantes**. 2020. Disponível em: https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enade.

INEP. **Microdados**. 2020. Disponível em: http://inep.gov.br/web/guest/microdados.

INEP. **Questionário do Estudante**. 2020. Disponível em: https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enade/questionario-do-estudante.

JIMÉNEZ-RAYGOZA, LI; MEDINA-VÁZQUEZ, AS; PÉREZ-TORRES, G. Proposal of a computer system for vocational guidance with data mining. *In*: IEEE. **2019 IEEE International Conference on Engineering Veracruz (ICEV)**. *[S.l.]*, 2019. v. 1, p. 1–5.

KEIM, Daniel A. Information visualization and visual data mining. **IEEE transactions on Visualization and Computer Graphics**, IEEE, v. 8, n. 1, p. 1–8, 2002.

KOHAVI, Ronny; QUINLAN, J Ross. Data mining tasks and methods: Classification: decision-tree discovery. *In*: **Handbook of data mining and knowledge discovery**. *[S.l.: s.n.]*, 2002. p. 267–276.

KURGAN, Lukasz A; MUSILEK, Petr. A survey of knowledge discovery and data mining process models. **The Knowledge Engineering Review**, Cambridge University Press, v. 21, n. 1, p. 1–24, 2006.

LAVRAČ, Nada. Selected techniques for data mining in medicine. **Artificial intelligence in medicine**, Elsevier, v. 16, n. 1, p. 3–23, 1999.

LI, Nan; COHEN, William W; KOEDINGER, Kenneth R; MATSUDA, Noboru. A machine learning approach for automatic student model discovery. *In*: ERIC. **Edm**. *[S.l.]*, 2011. p. 31–40.

LIU, Huan; MOTODA, Hiroshi. **Feature selection for knowledge discovery and data mining**. *[S.l.]*: Springer Science & Business Media, 2012. v. 454.

MACHADO, Sérgio P; FILHO, João Massena Melo; PINTO, Angelo C. A evasão nos cursos de graduação de química: uma experiência de sucesso feita no instituto de química da universidade federal do rio de janeiro para diminuir a evasão. **Química Nova**, SciELO Brasil, v. 28, p. S41–S43, 2005.

MACQUEEN, James *et al.* Some methods for classification and analysis of multivariate observations. *In*: OAKLAND, CA, USA. **Proceedings of the fifth Berkeley symposium on mathematical statistics and probability**. *[S.l.]*, 1967. v. 1, n. 14, p. 281–297.

MARTINS, Cleidis Beatriz Nogueira. Evasão de alunos nos cursos de graduação em uma instituição de ensino superior. **Montes Claros**, 2007.

MCGINNIS, William D; SIU, Chapman; ANDRE, S; HUANG, Hanyu. Category encoders: a scikit-learn-contrib package of transformers for encoding categorical data. **Journal of Open Source Software**, v. 3, n. 21, p. 501, 2018.

MICHIE, Donald; SPIEGELHALTER, David J; TAYLOR, CC *et al.* Machine learning. **Neural and Statistical Classification**, Technometrics, v. 13, n. 1994, p. 1–298, 1994.

MUCHERINO, Antonio; PAPAJORGJI, Petraq; PARDALOS, Panos M. **Data mining in agriculture**. *[S.l.]*: Springer Science & Business Media, 2009. v. 34.

NGAI, Eric WT; HU, Yong; WONG, Yiu Hing; CHEN, Yijun; SUN, Xin. The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. **Decision support systems**, Elsevier, v. 50, n. 3, p. 559–569, 2011.

NISBET, Robert; ELDER, John; MINER, Gary. **Handbook of statistical analysis and data mining applications**. *[S.l.]*: Academic Press, 2009.

OLIVEIRA, Bruna de; GUIMARÃES, Lucas José; SANTANA, Thainá Nunes Pires. O caminho para a redução da evasão de estudantes nas instituições de ensino superior. **Humanidades & Inovação**, v. 6, n. 18, p. 155–164, 2019.

PAREDES, Alberto Sànchez. **A evasão do terceiro grau em Curitiba**. *[S.l.]*: NUPES, 1994.

PAZ, Fábio; CAZELLA, Sílvio. Identificando o perfil de evasão de alunos de graduação através da mineração de dados educacionais: um estudo de caso de uma universidade comunitária. *In*: **Anais dos Workshops do Congresso Brasileiro de Informática na Educação**. *[S.l.: s.n.]*, 2017. v. 6, n. 1, p. 624.

PAZZANI, Michael; BILLSUS, Daniel. Learning and revising user profiles: The identification of interesting web sites. **Machine learning**, Springer, v. 27, n. 3, p. 313–331, 1997.

PLATT, John *et al.* Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. **Advances in large margin classifiers**, Cambridge, MA, v. 10, n. 3, p. 61–74, 1999.

PORTAL INEP. Disponível em: <http://portal.inep.gov.br/artigo/-/asset_publisher/B4AQV9zFY7Bv/content/baixa-ocupacao-de-vagas-remanescentes-revelada-pelo-censo-da-educacao-superior-inspira-nova-politica-do-mec-para-as-universidades-federais/21206>. **Acesso em 16 out. 2020**, 2018.

RAHMAH, Nadia; SITANGGANG, Imas Sukaesih. Determination of optimal epsilon (eps) value on dbscan algorithm to clustering data on peatland hotspots in sumatra. *In*: IOP PUBLISHING. **IOP conference series: earth and environmental science**. *[S.l.]*, 2016. v. 31, n. 1, p. 012012.

RASCHKA, Sebastian. **Python machine learning**. *[S.l.]*: Packt publishing ltd, 2015.

RESNICK, Paul; VARIAN, Hal R. Recommender systems. **Communications of the ACM**, ACM New York, NY, USA, v. 40, n. 3, p. 56–58, 1997.

RICCI, Francesco; ROKACH, Lior; SHAPIRA, Bracha. Introduction to recommender systems handbook. *In*: **Recommender systems handbook**. *[S.l.]*: Springer, 2011. p. 1–35.

RIGO, Sandro José; CAMBRUZZI, Wagner; BARBOSA, Jorge LV; CAZELLA, Sílvio C. Aplicações de mineração de dados educacionais e learning analytics com foco na evasão escolar: oportunidades e desafios. **Revista Brasileira de Informática na Educação**, v. 22, n. 01, p. 132, 2014.

ROMAO, Luiz. Análise do uso de técnicas de pré-processamento de dados em algoritmos para classificação de proteínas general terms. 05 2016.

ROMERO, C.; VENTURA, S. Educational data mining: A review of the state of the art. **IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)**, v. 40, n. 6, p. 601–618, 2010.

ROUSSEEUW, Peter J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. **Journal of computational and applied mathematics**, Elsevier, v. 20, p. 53–65, 1987.

RUSSELL, Stuart J; NORVIG, Peter. **Inteligência artificial**. *[S.l.]*: Elsevier, 2004.

SANDER, Jörg; ESTER, Martin; KRIEGEL, Hans-Peter; XU, Xiaowei. Density-based clustering in spatial databases: The algorithm gdbscan and its applications. **Data mining and knowledge discovery**, Springer, v. 2, n. 2, p. 169–194, 1998.

SCALI, Danyelle Freitas. **Evasão nos cursos Superiores de Tecnologia: a percepção dos estudantes sobre seus determinantes. 2009. 150 f**. 2009. Tese (Doutorado) — Dissertação (Mestrado em Educação)-Universidade Estadual de Campinas, Campinas, 2009.

SHALABI, Luai Al; SHAABAN, Zyad; KASASBEH, Basel. Data mining: A preprocessing engine. **Journal of Computer Science**, Citeseer, v. 2, n. 9, p. 735–739, 2006.

SILVA, Izaqueline Jhusmicele Alcântara da; MIRANDA, Gilberto José; LEAL, Edvalda Araujo; PEREIRA, Janser Moura. Estratégias das coordenações dos cursos de ciências contábeis para combater a evasão. **Revista Universo Contábil**, v. 14, n. 2, p. 61–81, 2019.

SINGHAL, Amit *et al.* Modern information retrieval: A brief overview. **IEEE Data Eng. Bull.**, v. 24, n. 4, p. 35–43, 2001.

TERVEEN, Loren; HILL, Will. Beyond recommender systems: Helping people help each other. **HCI in the New Millennium**, Citeseer, v. 1, n. 2001, p. 487–509, 2001.

TINTO, Vincent. **Leaving college: Rethinking the causes and cures of student attrition.** *[S.l.]*: ERIC, 1987.

TOSCHER, Andreas; JAHRER, Michael. Collaborative filtering applied to educational data mining. **KDD cup**, 2010.

VERMA, Manish; SRIVASTAVA, Mauly; CHACK, Neha; DISWAR, Atul Kumar; GUPTA, Nidhi. A comparative study of various clustering algorithms in data mining. **International Journal of Engineering Research and Applications (IJERA)**, v. 2, n. 3, p. 1379–1384, 2012.

WILLIAMSON, David F; PARKER, Robert A; KENDRICK, Juliette S. The box plot: a simple visual method to interpret data. **Annals of internal medicine**, American College of Physicians, v. 110, n. 11, p. 916–921, 1989.

WITTEN, Ian H; FRANK, Eibe. Data mining: practical machine learning tools and techniques with java implementations. **Acm Sigmod Record**, ACM New York, NY, USA, v. 31, n. 1, p. 76–77, 2002.

ZAKI, Mohammed J. Parallel and distributed data mining: An introduction. *In*: **Large-scale parallel data mining**. *[S.l.]*: Springer, 2000. p. 1–23.

# APÊNDICES

# APÊNDICE A – ATRIBUTOS CATEGÓRICOS APÓS LIMPEZA DE DADOS

Quadro de Atributos Categóricos

| Variável | Descrição | Informação Antes | Informação Depois |
|---|---|---|---|
| QE_I01 | Qual o seu estado civil? | Solteiro(a) = 71,23%<br>Casado(a) = 20,63%<br>Separado(a) = 3,18%<br>Outro = 2,81%<br>Viúvo(a) = 0,27%<br>Vazias = 1,88% | Solteiro(a) = 72,93%<br>Casado(a) = 20,78%<br>Outro = 6,29% |
| QE_I02 | Qual é a sua cor ou raça? | Branca = 52,17%<br>Parda = 32,36%<br>Preta = 8,73%<br>Amarela = 2,37%<br>N,Q,D, = 2,15%<br>Inválidas = 1,88%<br>Indígena = 0,34% | Branca = 53,29%<br>Parda = 32,94%<br>Preta = 8,86%<br>Outra = 4,91% |
| QE_I04 | Até que etapa de escolarização seu pai concluiu? | Ensino Médio = 30,93%<br>1° ao 5° ano = 26,26%<br>6° ao 9° ano = 14,80%<br>Ensino Sup, = 14,18%<br>Nenhuma = 6,83%<br>Pós-graduação = 5,12%<br>Inválidas = 1,88% | Ensino Médio = 31,50%<br>1° ao 5° ano = 26,78%<br>6° ao 9° ano = 15,10%<br>Ensino Sup, = 14,45%<br>Nenhuma = 6,93%<br>Pós-graduação = 5,24% |
| QE_I05 | Até que etapa de escolarização sua mãe concluiu? | Ensino Médio = 32,62%<br>1° ao 5° ano: = 21,87%<br>Ens, Superior = 15,62%<br>6° ao 9° ano = 14,36%<br>Pós-graduação = 8,95%<br>Nenhuma = 4,67%<br>Inválidas = 1,88% | Ensino Médio = 33,24%<br>1° ao 5° ano= 22,28%<br>Ensino Superior = 15,95%<br>6° ao 9° ano = 14,65%<br>Pós-graduação = 9,15 %<br>Nenhuma = 4,73% |
| QE_I06 | Com quem você mora atualmente? | Com parentes = 54,24%<br>Com cônjuge e/ou filhos= 27,73%<br>Sozinho = 9,08%<br>Com outras pessoas = 6,79%<br>Inválidas = 1,88%<br>Em alojamento da própria instituição = 0,28% | Com parentes = 55,74%<br>Com cônjuge e/ou filhos = 27,89%<br>Outros = 16,37% |
| QE_I07 | Quantas pessoas da sua família moram com você? Considere seus pais, irmãos, cônjuge, filhos e outros parentes que moram na mesma casa com você, | Três = 24,07%<br>Duas = 22,90%<br>Uma = 15,83%<br>Quatro = 13,48%<br>Nenhuma = 12,65%<br>Cinco = 5,63%<br>Seis = 2,09%<br>Inválidos = 1,88%<br>Sete ou mais = 1,47% | 3 = 25,32%<br>2 = 23,04%<br>1 = 15,27%<br>4 = 14,68%<br>0 = 11,25%<br>5 = 6,26%<br>6 = 2,38%<br>7 = 1,80% |
| QE_I08 | Qual a renda total de sua família, incluindo seus rendimentos? (em salários mínimos) | De 1,5 a 3 = 27,61%<br>De 3 a 4,5 = 20,50%<br>Até 1,5 = 19,73%<br>De 6 a 10 = 11,14%<br>De 4,5 a 6 = 10,86%<br>De 10 a 30 = 6,93%<br>Inválidos = 1,88%<br>Acima de 30 = 1,35% | Baixa = 48,22<br>Media = 43,33<br>Alta = 8,45 |

| | | | |
|---|---|---|---|
| QE_I09 | Qual alternativa a seguir melhor descreve sua situação financeira (incluindo bolsas)? | Não tenho renda e meus gastos são financiados pela minha família ou por outras pessoas = 27,64%<br>Tenho renda, mas recebo ajuda da família ou de outras pessoas para financiar meus gastos = 26,75%<br>Tenho renda e contribuo com o sustento da família = 17,61%<br>Tenho renda e não preciso de ajuda para financiar meus gastos = 10,43%<br>Sou o principal responsável pelo sustento da família = 8,27%<br>Não tenho renda e meus gastos são financiados por programas governamentais = 7,42%<br>Inválidas = 1,88% | Não tenho renda e meus gastos são financiados pela minha família ou por outras pessoas = 28,42%<br>Tenho renda, mas recebo ajuda da família ou de outras pessoas para financiar meus gastos = 27,42%<br>Tenho renda e contribuo com o sustento da família = 17,83%<br>Tenho renda e não preciso de ajuda para financiar meus gastos = 10,48%<br>Não tenho renda e meus gastos são financiados por programas governamentais = 8,26%<br>Sou o principal responsável pelo sustento da família = 7,59% |
| QE_I10 | Qual alternativa a seguir melhor descreve sua situação de trabalho (exceto estágio ou bolsas)? | Não estou trabalhando = 39,46%<br>40 horas semanais ou mais = 34,31%<br>De 21 a 39 horas semanais = 10,84%<br>Eventualmente = 7,12%<br>Até 20 horas semanais = 6,39%<br>Inválidas = 1,88% | 0 = 40,64%<br>1 = 34,61%<br>0,75 = 11,05%<br>0,25 = 7,18%<br>0,5 = 6,52% |
| QE_I11 | Que tipo de bolsa de estudos ou financiamento do curso você recebeu para custear todas ou a maior parte das mensalidades? No caso de haver mais de uma opção, marcar apenas a bolsa de maior duração, | Nenhum, embora meu curso não seja gratuito = 30,68%<br>Nenhum, pois meu curso é gratuito = 23,13%<br>FIES, apenas = 18,38%<br>Bolsa oferecida pela própria instituição = 8,29%<br>ProUni integral = 7,63%<br>Bolsa oferecida por governo estadual, distrital ou municipal = 2,48%<br>Bolsa oferecida por outra entidade (empresa, ONG, outra) = 2,46%<br>Inválida = 1,88%<br>ProUni parcial, apenas = 1,68%<br>Financiamento oferecido pela própria instituição = 1,67%<br>ProUni Parcial e FIES = 1,16%<br>Financiamento bancário = 0,56% | Prouni e/ou Fies (parcial ou integral) = 31,06%<br>Nenhum, embora meu curso não seja gratuito = 29,65%<br>Nenhum, pois meu curso é gratuito = 23,49%<br>Outro = 15,80% |
| QE_I16 | Em que unidade da Federação você concluiu o ensino médio? | São Paulo (SP) = 22,88%<br>Minas gerais (MG) = 11,83%<br>Rio de Janeiro (RJ) = 8,80%<br>Paraná (PR) = 6,40%<br>Rio Grande do Sul (RS) = 6,12%<br>Bahia (BA) = 5,18%<br>Santa Catarina (SC) = 4,17%<br>Pernambuco (PE) = 3,64%<br>Ceará (CE) = 3,46%<br>Goiás (GO) = 3,17%<br>Pará (PA) = 3,06%<br>Maranhão (MA) = 2,01%<br>Distrito federal (DF) = 1,97%<br>Espírito Santo (ES) = 1,95%<br>Paraíba (PB) = 1,78%<br>Amazonas (AM) = 1,73% | São Paulo (SP) = 22,96%<br>Minas gerais (MG) = 11,75%<br>Rio de Janeiro (RJ) = 8,73%<br>Paraná (PR) = 6,45%<br>Rio Grande do Sul (RS) = 6,19%<br>Bahia (BA) = 5,17%<br>Santa Catarina (SC) = 4,18%<br>Pernambuco (PE) = 3,64%<br>Ceará (CE) = 3,46%<br>Goiás (GO) = 3,19%<br>Pará (PA) = 3,07%<br>Maranhão (MA) = 2,02%<br>Distrito federal (DF) = 1,92%<br>Espírito Santo (ES) = 1,91%<br>Paraíba (PB) = 1,80%<br>Amazonas (AM) = 1,71% |

| | | | |
|---|---|---|---|
| | | Rio Grande do Norte (RN) = 1,50%<br>Mato Grosso (MT) = 1,47%<br>Piauí (PI) = 1,38%<br>Alagoas (AL) = 1,31%<br>Acre (AC) = 1,27%<br>Mato Grosso do Sul (MS) = 1,25%<br>Sergipe (SE) = 0,97%<br>Rondônia (RO) = 0,88%<br>Tocantins (TO) = 0,75%<br>Amapa (AP) = 0,42%<br>Não se aplica = 0,39%<br>Roraima (RR) = 0,26% | Rio Grande do Norte (RN) = 1,49%<br>Mato Grosso (MT) = 1,49%<br>Piauí (PI) = 1,39%<br>Alagoas (AL) = 1,28%<br>Acre (AC) = 1,27%<br>Mato Grosso do Sul (MS) = 1,25%<br>Sergipe (SE) = 0,98%<br>Rondônia (RO) = 0,90%<br>Tocantins (TO) = 0,76%<br>Amapa (AP) = 0,39%<br>Não se aplica = 0,39%<br>Roraima (RR) = 0,26% |
| QE_I17 | Em que tipo de escola você cursou o ensino médio? | Todo em escola pública = 64,28%<br>Todo em escola privada = 25,39%<br>Maior parte em escola pública = 4,37%<br>A maior parte em escola privada (particular) = 3,72%5<br>Inválidos = 1,88%<br>Parte no Brasil e parte no exterior = 0,25%<br>Todo no exterior = 0,11% | Maior parte/Toda em escola pública = 60,87%<br>Outros = 39,13% |
| QE_I19 | Quem mais lhe incentivou a cursar a graduação? | Pais = 61,11%<br>Ninguém = 17,52%<br>Outros membros da família que não os pais = 8,60%<br>Colegas/Amigos = 5,01%<br>Outras pessoas = 3,27%<br>Professores = 2,28%<br>Inválidos = 1,88%<br>Líder ou representante religioso = 0,40% | Pais = 62,56%<br>Outros = 19,79%<br>Ninguém = 17,65% |
| QE_I20 | Algum dos grupos abaixo foi determinante para você enfrentar dificuldades durante seu curso superior e conclui-lo? | Pais = 38,47%<br>Não tive dificuldade = 26,57%<br>Colegas de curso ou amigos = 10,11%<br>Outro grupo = 7,06%<br>Não recebi apoio = 5,96%<br>Professores do curso = 3,99%<br>Irmãos, primos ou tios = 2,36%<br>Inválidos = 1,88%<br>Avós = 1,48%<br>Colegas de trabalho = 1,33%<br>Líder ou representante religioso = 0,42%<br>Profissionais do serviço de apoio ao estudante da IES = 0,37% | Pais = 39,46%<br>Não tive dificuldade = 27,54%<br>Outros = 27,06%<br>Não recebi apoio = 5,94% |
| QE_I21 | Alguém em sua família concluiu um curso superior? | Sim = 68,55%<br>Não = 31,45% | Sim = 67,92%<br>Não = 32,08% |
| QE_I22 | Excetuando-se os livros indicados na bibliografia do seu curso, quantos livros você leu neste ano? | Um ou dois = 37,05%<br>De três a cinco = 28,65%<br>Nenhum = 14,01%<br>Mais de oito = 10,17%<br>De seis a oito = 8,24%<br>Inválidas = 1,88% | Um ou mais = 85,84%<br>Nenhum = 14,16% |
| QE_I23 | Quantas horas por semana, aproximadamente, você dedicou aos | Um ou dois = 41,98%<br>De três a cinco = 28,55%<br>De seis a oito = 11,67%<br>Mais de oito = 10,46% | 0,25 = 42,39%<br>0,5 = 29,20%<br>0,75 = 12,00%<br>1 = 10,76% |

| | | | |
|---|---|---|---|
| | estudos, excetuando as horas de aula? | Nenhum = 5,45%<br>Inválidas = 1,89% | 0 = 5,65% |
| QE_I24 | Você teve oportunidade de aprendizado de idioma estrangeiro na Instituição? | Não = 67,67%<br>Sim, somente na modalidade presencial = 18,37%<br>Sim, na modalidade a distância = 6,79%<br>Sim, parte na modalidade presencial e parte na modalidade semipresencial = 4,17%<br>Inválidas = 1,89%<br>Sim, somente na modalidade semipresencial = 1,11% | Não = 69,01%<br>Sim = 30,99% |
| QE_I25 | Qual o principal motivo para você ter escolhido este curso? | Vocação = 32,49%<br>Inserção no mercado de trabalho = 24,69%<br>Outro motivo = 15,93%5<br>Valorização profissional = 14,80%<br>Influência familiar = 6,82%<br>Inválidas = 1,88%<br>Oferecido na modalidade a distância = 1,48%<br>Prestígio Social = 1,12%<br>Baixa concorrência para ingresso = 0,79% | Vocação = 33,27%<br>Outros = 26,54%<br>Inserção no mercado de trabalho = 25,11%<br>Valorização profissional = 15,08% |
| QE_I26 | Qual a principal razão para você ter escolhido a sua instituição de educação superior? | Qualidade/reputação, 27,04%<br>Proximidade da minha residência = 20,64%<br>Gratuidade = 15,28%<br>Preço da mensalidade = 10,94%<br>Facilidade de acesso = 8,30%<br>Outro motivo = 7,03%<br>Possibilidade de ter bolsa de estudo = 4,86%<br>Inválidas = 2,29%<br>Proximidade do meu trabalho = 2,11%<br>Foi a única onde tive aprovação = 1,51% | Outro motivo = 38,12%<br>Qualidade/reputação = 22,74%<br>Proximidade da minha residência = 17,39%<br>Gratuidade = 12,70%<br>Preço da mensalidade = 9,05% |

# APÊNDICE B – DIAGRAMAS DE CAIXA E HISTOGRAMAS DE ATRIBUTOS

Neste capítulo estão presentes as Figuras que ilustram os Diagramas de Caixa e os Histogramas dos atributos *NU_IDADE*, *ANO_FIM_EM* e *ANO_IN_GRAD.*

Figura 1 - Diagrama de Caixa do atributo NU_IDADE antes e depois da eliminação das anomalias

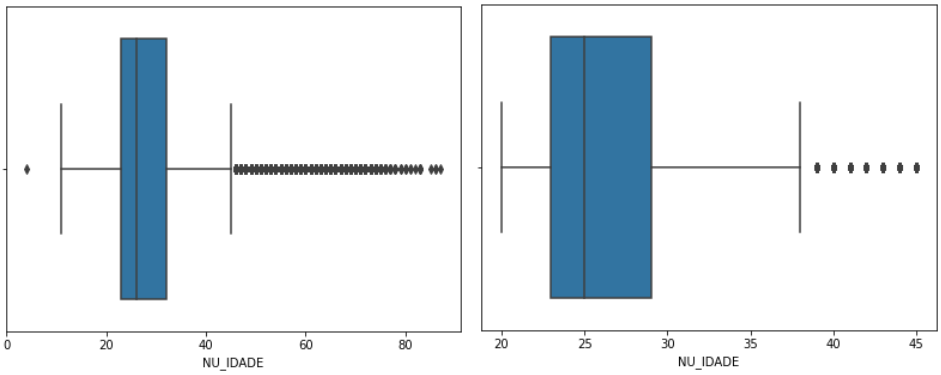



Figura 2 - Histograma do atributo NU_IDADE antes e depois da eliminação das anomalias

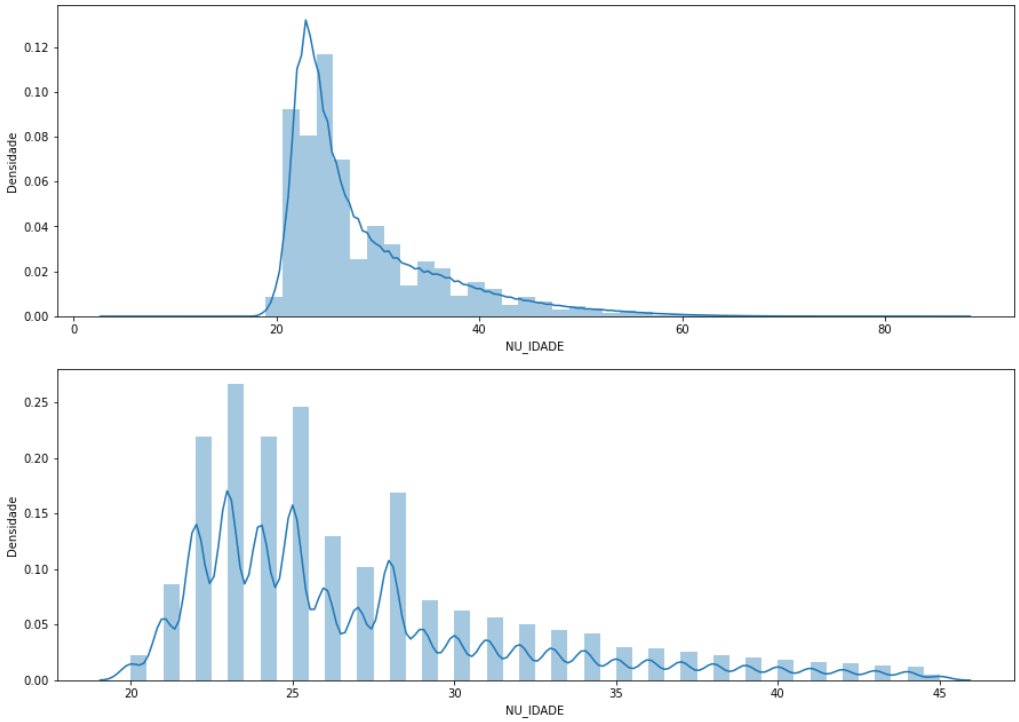


Figura 3 - Diagrama de Caixa do atributo ANO_FIM_EM antes e depois da eliminação das anomalias

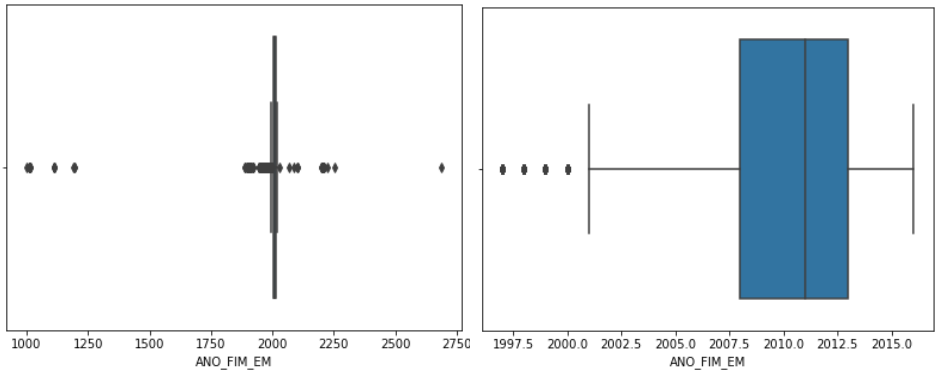



Figura 4 - Histograma do atributo ANO_FIM_EM antes e depois da eliminação das anomalias

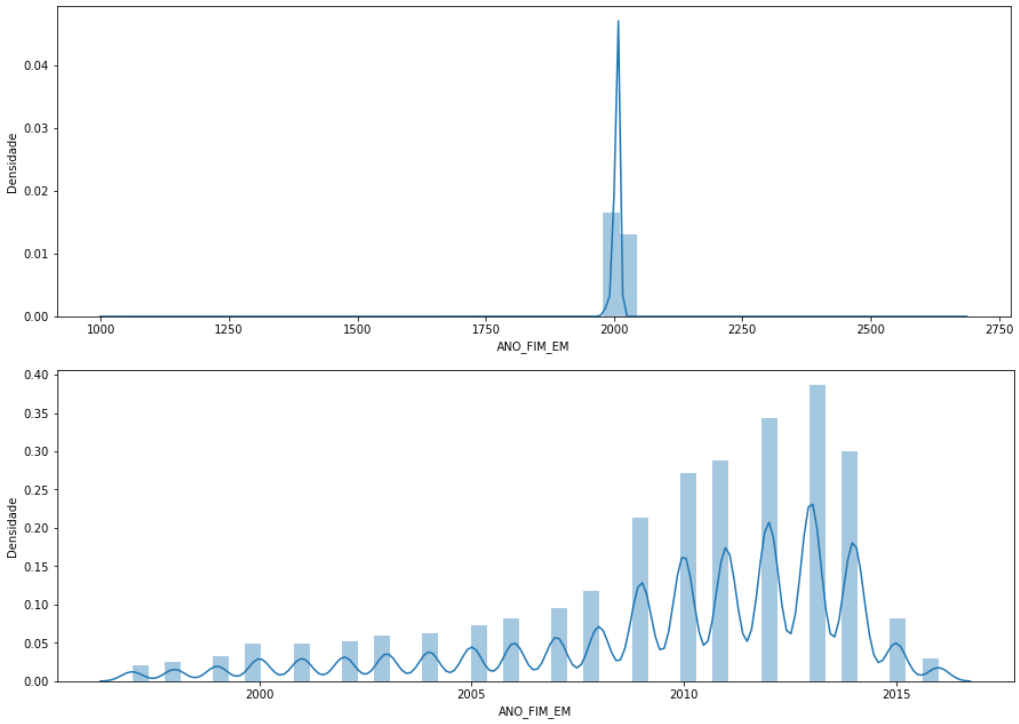


Figura 5 - Diagrama de Caixa do atributo ANO_IN_GRAD antes e depois da eliminação das anomalias

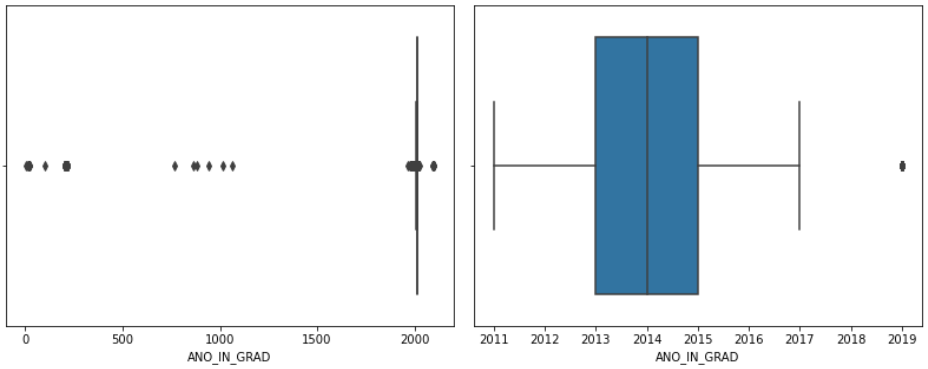
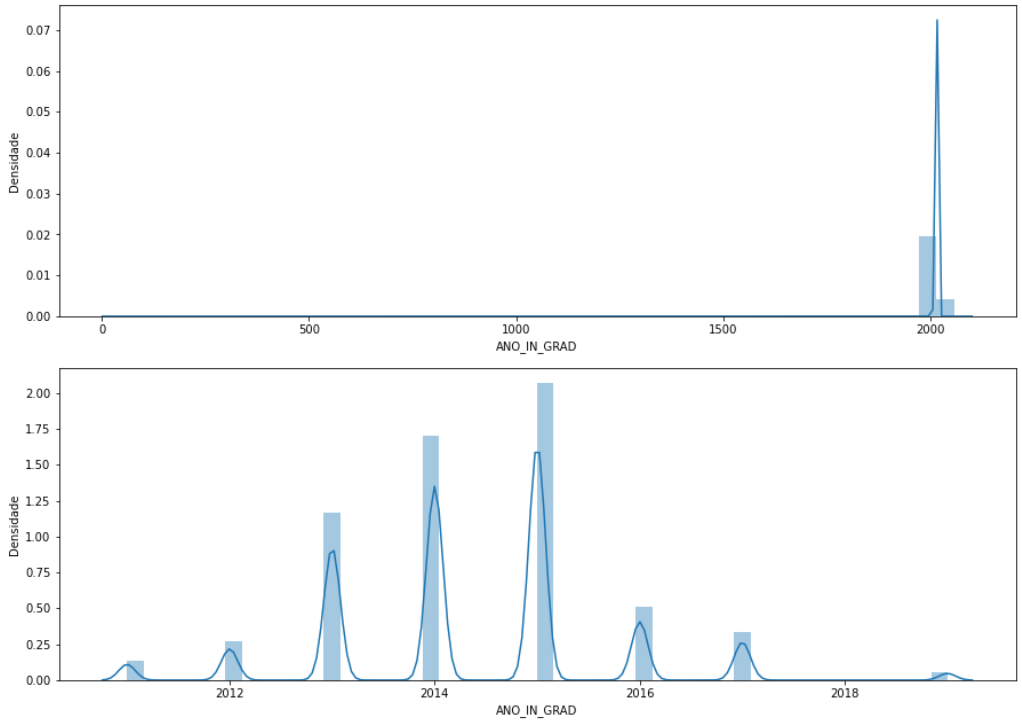



Figura 6 - Histograma do atributo ANO_IN_GRAD  antes e depois da eliminação das anomalias



# APÊNDICE C – QUADRO DE TAXAS DE ACERTO MÉDIAS PARA O ALGORITMO K-*MEANS*

| ÁREA DE CURSO | QTD DE AMOSTRAS | QTD DE UNIVERSIDADES | TAXA DE ACERTO MÉDIA (%) |
|---|---|---|---|
| **Pedagogia (Licenciatura)** | 111.267 | 709 | 8,64 |
| **Direito** | 126.627 | 941 | 12,39 |
| **Tecnologia Em Gestão De Recursos Humanos** | 18.727 | 305 | 14,24 |
| **Administração** | 98.900 | 1.172 | 14,55 |
| **Ciências Contábeis** | 52.610 | 729 | 14,69 |
| **Engenharia Civil** | 98.396 | 543 | 16,11 |
| **Tecnologia Em Logística** | 9.724 | 176 | 16,22 |
| **Engenharia De Produção** | 45.761 | 458 | 17,56 |
| **Tecnologia Em Análise e Desenvolvimento De Sistemas** | 9.637 | 198 | 17,63 |
| **Matemática (Licenciatura)** | 10.680 | 231 | 17,70 |
| **Letras-Português E Inglês (Licenciatura)** | 6.521 | 139 | 17,79 |
| **Psicologia** | 37.368 | 431 | 17,82 |
| **Educação Física (Bacharelado)** | 31.561 | 324 | 17,94 |
| **Enfermagem** | 38.173 | 549 | 18,32 |
| **Sistemas De Informação** | 11.888 | 291 | 18,96 |
| **Tecnologia Em Gestão Financeira** | 4.982 | 99 | 19,07 |
| **Comunicação Social - Publicidade E Propaganda** | 13.922 | 232 | 19,57 |
| **Tecnologia Em Marketing** | 5.193 | 101 | 19,93 |
| **História (Licenciatura)** | 14.111 | 190 | 20,08 |
| **Fisioterapia** | 21.491 | 357 | 20,32 |
| **Educação Física (Licenciatura)** | 27.221 | 346 | 20,61 |
| **Tecnologia Em Gestão Da Tec. Da Informação** | 3.206 | 59 | 20,71 |
| **Agronomia** | 13.618 | 247 | 20,73 |
| **Medicina** | 20.613 | 215 | 20,79 |
| **Tecnologia Em Gastronomia** | 4.562 | 92 | 20,99 |
| **Tecnologia Em Processos Gerenciais** | 8.496 | 121 | 21,02 |
| **Tecnologia Em Gestão Comercial** | 4.678 | 86 | 21,87 |
| **Ciências Biológicas (Licenciatura)** | 13.820 | 262 | 21,98 |
| **Engenharia Mecânica** | 34.945 | 273 | 22,15 |
| **Farmácia** | 16.584 | 292 | 22,38 |
| **Engenharia Elétrica** | 30.593 | 299 | 22,72 |
| **Arquitetura E Urbanismo** | 52.283 | 314 | 23,10 |
| **Geografia (Licenciatura)** | 8.782 | 136 | 23,15 |
| **Nutrição** | 20.385 | 311 | 23,94 |
| **Letras-Português (Licenciatura)** | 11.591 | 168 | 23,95 |
| **Tecnologia Em Estética E Cosmética** | 5.279 | 122 | 23,99 |
| **Química (Licenciatura)** | 4.785 | 141 | 24,21 |
| **Serviço Social** | 21.187 | 214 | 24,24 |
| **Comunicação Social - Jornalismo** | 9.737 | 192 | 24,61 |
| **Tecnologia Em Comércio Exterior** | 1.856 | 35 | 24,87 |
| **Biomedicina** | 10.356 | 205 | 25,13 |
| **Física (Licenciatura)** | 2.758 | 101 | 25,26 |
| **Tecnologia Em Design De Interiores** | 2.124 | 54 | 25,37 |

| | | | |
|---|---|---|---|
| **Medicina Veterinária** | 1.3726 | 185 | 25,63 |
| **Tecnologia Em Gestão Hospitalar** | 1.188 | 28 | 25,86 |
| **Tecnologia Em Redes De Computadores** | 2.491 | 80 | 26,29 |
| **Ciências Biológicas (Bacharelado)** | 6.566 | 161 | 26,48 |
| **Engenharia Ambiental** | 15.135 | 219 | 26,69 |
| **Teologia** | 3.744 | 56 | 26,97 |
| **Ciência Da Computação (Bacharelado)** | 8.398 | 190 | 27,00 |
| **Tecnologia Em Design De Moda** | 1.372 | 39 | 27,32 |
| **Odontologia** | 17.026 | 184 | 27,65 |
| **Ciências Econômicas** | 8.005 | 134 | 28,97 |
| **Tecnologia Em Gestão Ambiental** | 3.976 | 56 | 29,28 |
| **Engenharia De Controle E Automação** | 9.477 | 137 | 29,71 |
| **Tecnologia Em Design Gráfico** | 2.575 | 51 | 30,18 |
| **Química (Bacharelado)** | 2.890 | 76 | 30,48 |
| **Tecnologia Em Gestão Da Qualidade** | 1.276 | 31 | 30,62 |
| **Engenharia Química** | 13.685 | 136 | 30,82 |
| **Zootecnia** | 2.645 | 72 | 30,96 |
| **Engenharia** | 7.577 | 117 | 31,53 |
| **Tecnologia Em Gestão Pública** | 3.808 | 46 | 31,90 |
| **Relações Internacionais** | 4.873 | 77 | 31,93 |
| **Ciências Sociais (Licenciatura)** | 2.522 | 57 | 32,78 |
| **Artes Visuais (Licenciatura)** | 3.528 | 49 | 32,82 |
| **Filosofia (Licenciatura)** | 3.666 | 73 | 33,10 |
| **Engenharia Da Computação** | 7.615 | 126 | 34,31 |
| **Letras - Inglês** | 2.372 | 58 | 34,36 |
| **Tecnologia Em Agronegócios** | 1.479 | 39 | 34,63 |
| **Filosofia (Bacharelado)** | 1.166 | 35 | 36,54 |
| **Tecnologia Em Radiologia** | 2.574 | 66 | 37,05 |
| **Matemática (Bacharelado)** | 388 | 22 | 37,49 |
| **Geografia (Bacharelado)** | 2.038 | 49 | 37,57 |
| **Letras-Português E Espanhol (Licenciatura)** | 2.340 | 55 | 37,83 |
| **Física (Bacharelado)** | 711 | 31 | 37,96 |
| **Ciências Sociais (Bacharelado)** | 2.045 | 48 | 38,24 |
| **Engenharia De Alimentos** | 3.440 | 64 | 38,85 |
| **Design** | 5.764 | 91 | 39,54 |
| **Administração Pública** | 3.459 | 35 | 40,97 |
| **Ciência Da Computação (Licenciatura)** | 1.080 | 27 | 41,69 |
| **Música (Licenciatura)** | 2.748 | 55 | 41,73 |
| **Turismo** | 2.669 | 57 | 42,00 |
| **História (Bacharelado)** | 1.251 | 32 | 42,18 |
| **Tecnologia Em Segurança No Trabalho** | 513 | 12 | 43,57 |
| **Engenharia Florestal** | 3.872 | 56 | 44,94 |
| **Tecnologia Em Gestão Da Produção Industrial** | 1.527 | 29 | 48,25 |
| **Letras-Português (Bacharelado)** | 475 | 14 | 49,66 |
| **Secretariado Executivo** | 1.096 | 22 | 51,28 |
| **Fonoaudiologia** | 2.420 | 54 | 52,94 |