UNIVERSIDADE FEDERAL DE ITAJUBÁ

MINERAÇÃO DE DADOS NA IDENTIFICAÇÃO DE PERFIS DE SOLICITANTES DE AUXÍLIO ESTUDANTIL EM UMA INSTITUIÇÃO FEDERAL DE ENSINO SUPERIOR.

Guilherme Gonçalves Horta Jardim Bastos

Monografia realizada sob orientação do Prof. Rafael de Magalhães Dias Frinhani
e coorientação da Profa. Vanessa Cristina Oliveira De Souza

Itajubá

2021

UNIVERSIDADE FEDERAL DE ITAJUBÁ

MINERAÇÃO DE DADOS NA IDENTIFICAÇÃO DE PERFIS DE SOLICITANTES DE AUXÍLIO ESTUDANTIL EM UMA INSTITUIÇÃO FEDERAL DE ENSINO SUPERIOR.

Monografia apresentada como trabalho final de graduação, requisito parcial para obtenção do título de Bacharel em Ciência da Computação, sob orientação do Prof. Rafael de Magalhães Dias Frinhani e coorientação da Profa. Vanessa Cristina Oliveira De Souza.

Itajubá

2021

Resumo

Recentemente na história brasileira, políticas de acesso a universidade foram ampliadas, sobretudo para a população socioeconomicamente mais vulnerável, trazendo consigo a necessidade de uma maior efetividade por parte dos programas de assistência estudantil. Falta aos gestores acadêmicos ferramentas que auxiliem o gerenciamento de programas e políticas de gestão acadêmica. Este trabalho busca construir e avaliar modelos de Mineração de Dados Educacionais para identificação de grupos de requerentes de assistência estudantil com perfil socioeconômico semelhantes, bem como a classificação de novos requerentes quanto a um determinado perfil. Diversos métodos foram executados e avaliados sendo que os modelos baseados nos algoritmos K-Means e K-Prototypes mostram os melhores resultados. Os resultados foram promissores mediante a descoberta de grupos com diferentes perfis, entretanto, não foi possível apontar um modelo ideal para tal, necessitando assim de pesquisas complementares para melhorar sua aplicação.

Palavras-chaves: Mineração de Dados Educacionais; Mineração de Dados Socioeconômicos; Aprendizado Não-Supervisionado.

Abstract

Recently, in brazilian history, democratization of access to university education were expanded, specially for the socioeconomically vulnerable population. This situation started a need for more effective programs of student assistance, and also academic managers lack tools to help them manage those programs and academic policies. This thesis tried to build and avaluate an Educational Data Mining model to classify groups with different profiles, applied to socioeconomic data of student assistance applicants in a federal institution of higher education. Several methods were executed and evaluated, and two models demonstrated to be promising, they were K-Means and K-Prototypes. The Data Mining results were optimistic on discovering groups with different profiles, however, it was not possible to point out an ideal model for such, thus requiring further research.

Keywords: Educational Data Mining; Data Mining on Socioeconomic Data; Unsupervised Learning.

Lista de Ilustrações

Lista de Tabelas

Lista de Abreviaturas e Siglas

CRISP-DM	*Cross Industry Standard Process for Data Mining*
DCBD	Descoberta de Conhecimento em Bancos de Dados
DM	*Data Mining*
DAE	Diretoria de Assuntos Estudantis
EDS	*Early Detection System*
EDM	*Educational Data Mining*
GPP	Gerência de Portfólio de Projetos
IFES	Instituições Federais de Ensino Superior
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
KDD	*Knowledge Discovery in Databases*
MEC	Ministério da Educação
MD	Mineração de Dados
MDE	Mineração de Dados Educacionais
SDP	Sistema de Detecção Precoce
Unicamp	Universidade Estadual de Campinas
UNIFEI	Universidade Federal de Itajubá
CEPEAD	Conselho de Ensino, Pesquisa, Extensão e Administração

Sumário

1 Introdução

No âmbito do ensino superior no Brasil, a política de assistência estudantil tem como objetivo a igualdade de oportunidades para os discentes, de modo a contribuir para a melhoria do desempenho acadêmico e agir, preventivamente, nas situações de repetência e evasão decorrentes da condição financeira (BRASIL, 2007). Entretanto, na perspectiva de Lobo et al. (2007), no Brasil faltam iniciativas, pesquisas e estudos sistemáticos sobre a evasão no sistema de ensino superior que indiquem com precisão quais são as melhores práticas para combatê-la. Para Nogueira (2017), um dos aspectos mais relevantes para este quadro é a questão socioeconômica do discente, que afeta sua saúde mental, seu rendimento acadêmico e, consequentemente, sua permanência no ambiente universitário.

Com base em uma amostra de discentes da Universidade Estadual de Campinas (Unicamp), Neves e Dalgalarrondo (2007) relatam que 58% dos alunos apresentaram algum tipo de sofrimento ou transtorno mental que influenciam no seu rendimento acadêmico, sendo que uma das principais causas é sua baixa condição socioeconômica. Estudos de 2004 e 2005 mostram que a média nacional de casos similares está entre 25% e 34%, dado que se mostra preocupante não apenas para a administração institucional, mas também para toda a sociedade (GIGLIO, 1976; CERCHIARI, 2004; FACUNDES; LUDERMIR, 2005). Dessa maneira, políticas de assistência estudantil são fundamentais para garantir a igualdade de oportunidades e a plena participação dos estudantes nas atividades regulares de seus respectivos cursos.

Algumas políticas públicas já foram implementadas nas Instituições Federais de Ensino Superior (IFES) na tentativa de mitigar a evasão por motivos socioeconômicos. Conforme descrito no parágrafo 1º do Artigo 3º da Portaria Normativa do Ministério da Educação (MEC) nº 39/2007 (BRASIL, 2007), “As ações de assistência estudantil devem considerar a necessidade de viabilizar a igualdade de oportunidades, contribuir para a melhoria do desempenho acadêmico e agir, preventivamente, nas situações de repetência e evasão decorrentes da insuficiência de condições financeiras". O parágrafo 2º do Artigo 3º do decreto nº 7234 (BRASIL, 2010) sobre o Programa Nacional de Assistência Estudantil (PNAES) descreve: “Caberá à instituição federal de ensino superior definir os critérios e a metodologia de seleção dos alunos de graduação a serem beneficiados”.

Um dos grandes desafios das IFES é o gerenciamento de tais políticas de forma eficiente, visto que a evasão e perda do desempenho acadêmico prejudicam a eficiência institucional, uma vez que o investimento público não teve o devido retorno. Na Universidade Federal de Itajubá (UNIFEI), as ações desse programa são executadas pela Diretoria de Assuntos Estudantis (DAE). Os critérios de seleção dos estudantes para as políticas de assistência estudantil levam em conta o perfil socioeconômico dos alunos e outros requisitos específicos, formulados

de acordo com a realidade da instituição. Estes critérios são definidos pela NORMA 2.3.03 (UNIFEI, 2020), e são avaliados pelo grupo técnico da DAE, composto por assistentes sociais, psicólogas e auxiliares.

A DAE precisa gerir recursos escassos para uma grande quantidade de discentes com necessidade financeira. Isso gera uma demanda considerável à equipe, que precisa analisar manualmente e de forma minuciosa todas as requisições de auxílio solicitadas. A demanda por auxílio vem crescendo ano após ano, ao contrário dos recursos, implicando em decisões e escolhas cada vez mais críticas.

Para Silva e Marques (2017), falta ao gestor ferramentas que possibilitem um entendimento maior sobre o problema enfrentado pelos discentes, de modo a aumentar as chances para seu enfrentamento a partir do contexto socioeconômico da IFES. Estudos têm sido desenvolvidos para descobrir meios de mitigar a evasão de alunos no ensino superior. Tontini e Walter (2014) constatam que técnicas de rede neural e análise de cluster são efetivas na identificação de prováveis evasores. Baseados em técnicas de *Educational Data Mining* (EDM), que é o campo de pesquisa que explora dados provenientes de ambientes educacionais, os estudos de Alencar et al. (2015) e Prado et al. (2011) apresentam resultados relevantes, com acurácia de até 91% na predição. Para Prado et al. (2011), os resultados da pesquisa indicam que é possível identificar alunos com risco elevado de evasão por meio de métodos de EDM.

Porém, estes estudos são realizados sobre dados do rendimento acadêmico dos discentes. Sendo assim, a proposta deste trabalho é realizar estudos a partir dos dados socioeconômicos dos solicitantes de assistência estudantil, para auxiliar a tomada de decisão no âmbito da gestão acadêmica. Parte-se da hipótese de que um modelo de classificação baseado em métodos de aprendizado de máquina pode contribuir para a identificação de grupos de solicitantes de assistência estudantil com diferentes perfis, e a classificação automática de novos solicitantes.

1.1 Motivação

Esta pesquisa é motivada principalmente por abordar um tópico importante para a manutenção de diversos discentes no ensino superior público do Brasil, sobretudo os oriundos de classes sociais menos favorecidas.

Além disto, auxiliar a equipe da DAE e os gestores acadêmicos da UNIFEI no processo de escolha dos bolsistas, desenvolvendo uma ferramenta com fundamentos em medidas matemáticas e estatísticas, reduzindo assim o aspecto discricionário do processo de seleção e acelerando uma tarefa que é realizada manualmente.

1.2 Objetivos

O objetivo deste trabalho é a identificação de perfis de solicitantes de assistência estudantil para auxiliar a tomada de decisão na concessão de bolsas. Para isso, serão desenvolvidos modelos de classificação baseados em métodos de aprendizado de máquina não-supervisionados e supervisionados para Mineração de Dados Educacionais. O trabalho tem os seguintes objetivos específicos:

- Propor um modelo de classificação para identificação de grupos com diferentes perfis socioeconômicos.
- Implementar e avaliar o desempenho de métodos de aprendizado não-supervisionados e supervisionados em relação a medidas de qualidade.
- Executar e comparar métodos de classificação que utilizem exclusivamente dados numéricos ou dados mistos (*mixed data*).

A pesquisa realizada neste trabalho tem a finalidade de contribuir com a discussão dos temas de Mineração de Dados Educacionais (EDM) e Mineração de Dados Socioeconômicos. Também é objetivo deste trabalho a implementação e validação dos modelos desenvolvidos, a partir da análise de qualidade da solução e desempenho dos métodos de aprendizado adotados. A validação visa verificar a viabilidade da aplicação do modelo, para auxiliar e acelerar o processo de seleção dos solicitantes a serem contemplados com a bolsa auxílio.

1.3 Contribuições

O trabalho contribui para o campo de pesquisa em EDM e para mineração de dados socioeconômicos. É também relevante no âmbito da gestão acadêmica, uma vez que pretende elaborar uma ferramenta de auxílio a gestão, que pode ser replicada no contexto de outras IFES.

1.4 Organização do Trabalho

Este documento se organiza no sentido de seguir as regras inerentes a uma monografia e de dispor seu conteúdo em um linha de raciocínio coerente para o leitor. Sendo assim, o Capítulo 1, antes apresentado, traz o contexto, a motivação e os objetivos envolvidos na pesquisa. O Capítulo 2 apresenta os principais conceitos e trabalhos relacionados ao tema. O Capítulo 3 discorre sobre os materiais e os métodos utilizados. O Capítulo 4 expõe as tarefas realizadas em cada etapa do desenvolvimento do trabalho. O Capítulo 5 ilustra os resultados obtidos e as discussões sobre tais. E as conclusões e trabalhos futuros são apresentadas no Capítulo 6.

2 Referencial Teórico

Com o advento cada vez mais célere de novas tecnologias eletrônicas, a quantidade de dados gerados atinge volumes de difícil análise por humanos. Por consequência, vêm sendo desenvolvidas formas de descobrir informações ocultas nestes dados para um melhor entendimento para a tomada de decisões.

As seções seguintes apresentam conceitos de Mineração de Dados (MD), mais especificamente a Mineração de Dados Educationais (MDE), que é uma das possibilidades para descoberta de conhecimento no auxílio a tomada de decisões.

2.1 Processo de Descoberta de Conhecimento

Como explica Fayyad et al. (1996), a descoberta de conhecimento em dados já foi tratada por diferentes nomes, como: extração de conhecimento, descoberta de informação, arqueologia de dados. Porém, em 1989 foi cunhado o termo *Knowledge Discovery in Databases* (KDD), ou em português Descoberta de Conhecimento em Bancos de Dados (DCBD).

Para Morais e Ambrósio (2007), o principal objetivo desta área está relacionado à descoberta de co-relacionamentos e dados implícitos em registros de bancos de dados, através do estudo e desenvolvimento de processos de extração de conhecimento. A intenção é encontrar conhecimento a partir de um conjunto de dados para ser utilizado em um processo decisório.

As etapas do processo de KDD são descritas por Fayyad et al. (1996) como: entendimento e seleção dos dados, pré-processamento e transformação dos dados, aplicação de técnicas de MD e por fim interpretação e avaliação dos padrões. Estas etapas são na visão de Silva et al. (2016) representadas de forma mais resumida pelo esquema da Figura 1. Para ele, o processo se inicia com a obtenção e organização da base de dados que se deseja descobrir um conhecimento útil, seguido pelo pré-processamento que consiste na alteração e exclusão de dados repetidos ou discrepantes e na seleção e normalização dos atributos mais relevantes. Com os dados pré-processados é executada a tarefa de mineração de dados propriamente dita, com seus diferentes métodos. E por fim os resultados são validados e avaliados em gráficos, tabelas e relatórios.

Figura 1: Etapas do processo de KDD, adaptado de Silva et al. (2016).

Goncalves e Freitas (2002) afirmam que a etapa de mineração é tão importante, que o termo "Mineração de Dados", ou do inglês *Data Mining* (DM), tem sido utilizado para identificar todo o processo, como um sinônimo para o processo de DCBD. A seção seguinte apresenta a MD segundo a literatura.

2.2 Mineração de Dados

A Mineração de Dados é a etapa processo de KDD cujo objetivo é de descobrir padrões interessantes e conhecimentos em uma grande quantidade de dados. Na MD busca-se obter informações de um conjunto de dados e convertê-las em um formato compreensível para uso adicional. Técnicas de mineração de dados são usadas para operar em grandes volumes de dados, para descobrir informações ocultas, padrões e relacionamentos úteis na tomada de decisão (HAN et al., 2011; MAJEED; NAAZ, 2018; BARADWAJ; PAL, 2011). Para Silva et al. (2016), trata-se da aplicação de técnicas, implementadas por meio de algoritmos computacionais, capazes de receber, como entrada, um conjunto de fatos ocorridos no mundo real e devolver, como saída, um padrão de comportamento, o qual pode ser expresso, por exemplo, como uma regra de associação, uma função de mapeamento ou a modelagem de um perfil.

2.2.1 Funcionalidades

A literatura apresenta diferentes formas de descrever e caracterizar as funcionalidades da mineração de dados. Segundo Fayyad et al. (1996), os dois principais objetivos da mineração de dados tendem a ser predição e descrição. A predição envolve o uso de algumas variáveis ou campos do banco de dados para prever valores desconhecidos ou valores futuros de outras variáveis de interesse. Já a descrição se concentra em encontrar padrões que descrevam os dados e sejam interpretáveis pelo homem. Essas tarefas podem ser especializadas. No caso da previsão, a classificação que trata de prever um rótulo, e a regressão que prevê uma quantidade. Já na descrição o agrupamento, a sumarização, a modelagem de dependências e detecção de desvios. No entanto, para Han et al. (2011), em um nível mais especializado, a tarefa de descrição inclui a mineração de padrões frequentes, associações e correlações; análise de grupos; e análise de *outliers*.

As funcionalidades em mineração de dados não são um consenso na literatura atual. No entanto, definir de forma clara o conceito da funcionalidade e a que resultado se pretende alcançar, é fundamental para o processo como um todo. Cortes et al. (2002) interpretam a MD conforme a Figura 2 (CORTES et al., 2002).

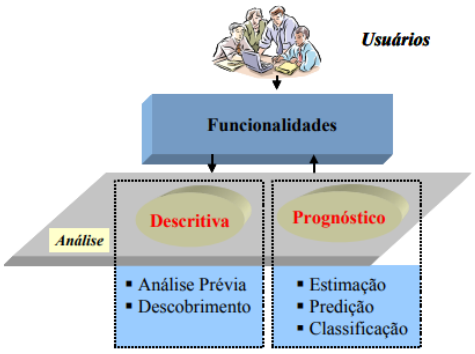


Figura 2: Funcionalidades em mineração de dados, adaptado de Cortes et al. (2002).

A Análise Descritiva representa a área de investigação nos dados que busca tanto descrever fatos relevantes, não-triviais e desconhecidos dos usuários, como analisar a base de dados, principalmente pelo seu aspecto de qualidade, para validar todo o processo da mineração e seus resultados (CORTES et al., 2002). Esta área é subdividida em duas funcionalidades:

- Análise Prévia: é o processo de analisar uma base de dados com o objetivo de identificar anomalias ou resultados raros que possam influenciar os resultados da mineração de dados.
- Descobrimento: é o processo de examinar uma base de dados com o objetivo de encontrar padrões escondidos, sem que necessariamente exista uma idéia ou hipótese clara previamente estabelecida.

Para cada funcionalidade existem diferentes técnicas, com propósitos específicos. A Figura 3 ilustra a relação entre as funcionalidades da Análise Descritiva e suas respectivas técnicas. (CORTES et al., 2002).

Já a Análise de Prognóstico, tratada por alguns autores apenas como predição (HAN et al., 2011; FAYYAD et al., 1996), busca inferir resultados com base nos padrões encontrados na Análise Descritiva. Cortes et al. (2002), se basendo principalmente na obra de Weiss e Indurkhya (1997), descreve suas subfuncionalidades de tal forma:

- Estimação: é o processo de predizer algum valor, baseado num padrão já conhecido. Por exemplo, conhecendo-se o padrão de despesas e a idade de uma pessoa, estimar se seu salário e número de filhos.

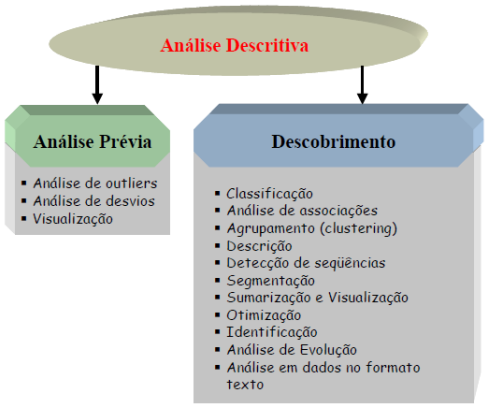


Figura 3: Lista de técnicas das funcionalidades Análise Prévia e Descobrimento da Análise Descritiva (CORTES et al., 2002).

- Predição: é o processo de predizer um comportamento futuro, baseado em vários valores. Por exemplo, baseado na formação escolar, no trabalho atual e no ramo de atividade profissional de uma pessoa, predizer que seu salário será de um certo montante até um determinado ano.

- Classificação: é o processo para predizer algum valor para uma variável categórica. Por exemplo, podemos num banco financeiro, determinar um conjunto de clientes que oferecem risco ou não para contrair um empréstimo pessoal.

2.2.2 Abordagens

As abordagens definem como o usuário irá realizar o processo de mineração de dados, e isso dependerá de seu objetivo final. Fundamentalmente existem as abordagens *top-down* e *bottom-up*. Na abordagem *top-down*, ou descrita por alguns autores como teste de hipótese, o usuário parte do princípio que existe uma hipótese ou uma idéia já concebida e deseja confirmá-la ou refutá-la. Na abordagem *bottom-up*, ou busca de conhecimento, o usuário explora os dados com o objetivo de descobrir alguma informação ainda não conhecida (BERRY; LINOFF, 2004; THURAISINGHAM, 1999). Estes dois conceitos são bem descritos por Cortes et al. (2002) na Tabela 1.

Tabela 1: Abordagens para Mineração de Dados (CORTES et al., 2002).

Mineração de Dados Abordagens/Metodologias	
Top-Down	**Botton-Up**
Inicia com a hipótese e valida a hipótese.	Analisa os dados e extrai padrões.
Hipóteses podem ser fornecidas inicialmente da abordagem Botton-Up ou de um conhecimento do mundo real.	Procura por alguma idéia pré-concebida (aplicação direta).
Se a hipótese não é satisfeita, então revisa-se a hipótese.	Não tem idéia do que está procurando (aplicação indireta).

2.3 Mineração de Dados Educacionais

Com a expansão dos cursos a distância e também daqueles com suporte computacional, muitos pesquisadores têm mostrado interesse em utilizar mineração de dados para investigar questões científicas na área de educação. Dentro deste contexto, surgiu uma área de pesquisa conhecida como Mineração de Dados Educacionais (BAKER et al., 2011). Segundo o site www.educationaldatamining.org, mantido pela Sociedade Internacional de Mineração de Dados Educacionais (*International Educational Data Mining Society*): "*A EDM é uma disciplina, preocupada em desenvolver métodos para explorar os tipos únicos de dados que provêm de ambientes educacionais e usar esses métodos para entender melhor os alunos, e as condições em que eles aprendem.*".

As instituições podem implementar decisões tomadas a partir da análise de dados, para ajudar a promover o acesso dos alunos, diminuir as taxas de retenção e orientar os programas de intervenção (PELAEZ et al., 2019). Para Baker et al. (2011), por meio da EDM é possível compreender de forma mais eficaz e adequada os alunos, como eles aprendem, o papel do contexto na qual a aprendizagem ocorre, além de outros fatores que influenciam a aprendizagem. Por exemplo, é possível identificar em que situação um tipo de abordagem instrucional proporciona melhores benefícios educacionais ao aluno.

2.3.1 Métodos

Métodos educacionais de mineração de dados geralmente são diferentes dos métodos padrões de mineração de dados, devido à necessidade de explicar explicitamente a hierarquia multinível e a não independência dos dados educacionais (BAKER, in press). A seguir são apresentadas as principais tarefas e algoritmos de EDM, que constam na categorização das subáreas de EDM, na taxonomia proposta por Baker et al. (2011).

- **Predição:** Nesta tarefa, a meta é desenvolver modelos que façam inferência sobre aspectos específicos dos dados por meio da análise e associação dos diversos aspectos encontrados nos dados (COSTA et al., 2012).

  Exemplo de algoritmos: Árvore de Decisão; Máquina de Vetores de Suporte; Regressão Linear.

- **Agrupamento:** O objetivo é dividir o conjunto de dados em grupos, de forma que os objetos, representados pelos dados, fiquem agrupados de acordo com a semelhança entre eles (BAKER et al., 2011).

  Exemplo de algoritmos: K-Means; Algoritmo Genético.

- **Mineração de Relações:** Nesta tarefa, o objetivo é descobrir possíveis relações entre as variáveis de um banco de dados. Isto pode ser feito investigando quais variáveis estão mais fortemente relacionadas com uma determinada variável de interesse, ou pela identificação de relações fortes entre quaisquer duas variáveis (BAKER et al., 2011).

Os métodos também podem ser divididos em não-supervisionados e supervisionados, onde a diferente principal é a existência de rótulos no subconjunto de dados de treinamento (BERRY et al., 2020). Estes métodos podem ser combinados de acordo com o objetivo da aplicação, tendo como exemplo este trabalho, onde um método não-supervisionado agrupa o conjunto de dados em diferentes *clusters*, e em seguida o método supervisionado determina as regras implícitas das características de cada subconjunto e a classificação de novos indivíduos.

Existem diferentes formas de executar o processo de agrupamento dos indivíduos, dentre os paradigmas de agrupamento, os dois principais são Hierárquico e Particional. O paradigma hierárquico inicializa todos os objetos como um *cluster* único, e a partir disso calcula a proximidade entre cada *cluster*, mesclando cada par de *clusters* entre os mais próximos e calculando novamente a nova proximidade entre os *clusters* restantes. Este processo é repetido até que se haja a quantidade de grupos desejada. Já nos algoritmos Particionais os grupos são construídos a partir de centróides, ou seja, objetos que representam o centro de um grupo. Os demais objetos são agrupados a partir da proximidade a estes centróide. Este processo é calculado a cada iteração do algoritmo, e para cada iteração, a distância do centroide para cada *cluster* é atualizada. Ao fim, os melhores resultados são apresentados. O trabalho de Reynolds et al. (2006) realiza uma comparação entre estes dois paradigmas.

Estas distâncias e proximidades são calculas a partir de um Medidas de Distância, que são utilizadas para definir um valor para o distanciamento entre dois objetos. Entre elas, as medidas Euclidiana e Gower, que são utilizadas neste trabalho, são resumidas a seguir:

- **Euclidiana:** esta medida calcula o comprimento de um segmento de linha entre dois pontos no espaço euclidiano ou coordenadas cartesianas. Suas características são descritas detalhadamente no livro de Deza e Deza (2009), e sua fórmula é expressa por:

$$E_{(x,y)} = \sqrt{\sum_{i=o}^{n} (x_i - y_i)^2} \tag{2.1}$$

  onde x é o valor do primeiro objeto em uma das dimensões do espaço euclidiano (lê-se cada atributo do conjunto de dados); y é o valor do segundo objeto na mesma dimensão do espaço euclidiano; e n é o número de dimensões do espaço euclidiano (quantidade de atributos do conjunto de dados).

- **Gower:** este coeficiente é usado para calcular a distância entre duas entidades cujos atributos tem uma combinação de valores categóricos e numéricos, gerando uma matriz de distância que pode ser consumida por um método de agrupamento. Este coeficiente foi proposto por Gower (1971) e é expresso pela a fórmula a seguir:

$$S_{ij} = \frac{\sum_{i=1}^{p} W_{ijk}.S_{ijk}}{\sum_{i=1}^{p} W_{ijk}} \tag{2.2}$$

  onde k é o número de variáveis, variando de 1 a p; i e j são dois registros quaisquer; W_{ijk} é o peso dado a comparação ijk, atribuindo valor 1 para comparações válidas e valor 0 para comparações onde a variável está ausente em um ou ambos indivíduos; S_{ijk} é a contribuição da variável k na similaridade entre os indivíduos i e j.

Os algoritmos de métodos não-supervisionados e supervisionados utilizados neste trabalho são brevemente descritos a seguir, com a identificação de seus paradigmas e as medidas de distância utilizadas:

- **Não-Supervisionados**

  1. **K-Means:** este é um algoritmo particional e tem como objetivo descobrir k partições no conjunto de dados que sejam disjuntas entre si. Seu funcionamento é de maneira iterativa variando os centroides, para encontrar o melhor centroide para cada grupo. Assim o objeto estará associado ao grupo cujo centroide se encontra mais próximo a partir da medida euclidiana. (SILVA et al., 2016).
  2. **Agglomerative Clustering:** este algoritmo constrói uma estrutura hierárquica, dividindo os *clusters* recursivamente de baixo para cima. Cada objeto representa inicialmente um cluster próprio, que são fundidos sucessivamente a partir da proximidade euclidiana até atigir a estrutura de *cluster* desejada (MAIMON; ROKACH, 2010).

3. **Birch:** este é um algoritmo hierárquico que geralmente é utilizado em grandes conjuntos de dados, principalmente por conta de sua performance. Seu procedimento envolve a criação de um resumo pequeno e compacto do conjunto de dados que retém o máximo de informação possível, só então sua divisão em grupos é realizada de forma incremental e dinâmica (ZHANG et al., 1996; SILVA et al., 2016). É uma ótima opção para se utilizar em dados com muitos *outliers*. E foi calculada a partir da medida euclidiana.
4. **K-Medoids:** este é um algoritmo particional muito similar ao K-Means, diferindo principalmente na forma em que os centroides são encontrados (SHENG; LIU, 2004; BRITO et al., 2011). Sua execução se deu sobre a matriz de distância gerada pelo coeficiência de distância Gower.
5. **K-Prototypes:** este é um algoritmo particional desenvolvido para permitir o agrupamento de objetos com atributos numéricos e categóricos. Por meio da definição de uma medida de dissimilaridade combinada, o algoritmo incorpora o resultado do conjunto numérico processado pelo K-Means e o resultado do conjunto categórico processado pelo K-Modes (HUANG, 1998). E a medida de distância utilizada foi a euclidiana.

- **Supervisionados**

1. **Árvore de Decisão C4.5:** este algoritmo transforma um fato implícito ao conjunto de dados e seus subconjuntos, em uma árvore de decisão que representa as regras dessa divisão. As regras podem ser facilmente entendidas com linguagem natural. Seu principal benefício é a sua capacidade de tornar processos complexos de tomada de decisão em processos mais simples (DAMANIK et al., 2019).

2.3.2 Aplicações

Em estudo recente, Berens et al. (2019) procuram desenvolver um Sistema de Detecção Precoce (SDP), do inglês *Early Detection System* (EDS), para prever a evasão de alunos, por meio de método de aprendizado de máquina adaptado a EDM, usando dados administrativos de estudantes de universidades estaduais e privadas. Os autores chegaram a resultados satisfatórios, com precisão das previsões no final do primeiro semestre de 79% para a universidades estaduais e 85% para a universidade privadas de ciências aplicadas. Após o quarto semestre, a precisão aumenta para 90% para as universidades estaduais e 95% para as universidades privadas de ciências aplicadas.

Já Pelaez et al. (2019), utilizando técnicas de classificação e árvore de decisão em dados do curso de psicologia na Universidade Estadual de San Diego, identificam três subgrupos de alunos com risco de evasão. Os alunos que foram identificados nos grupos de menor risco são principalmente estudantes que moram no campus e se identificam como brancos. Nos dois

grupos de baixo risco explorados no trabalho, havia poucos alunos cujos pais não tinham ensino superior. E os dois grupos com maior risco representam estudantes provenientes de populações historicamente sub-representadas no ensino superior, em sua maioria mexico-americanos.

Utilizando-se de bases de dados educacionais fornecidas pelo INEP e aplicando técnicas de mineração de dados, Nascimento et al. (2018) apresentam um estudo com a finalidade de melhor explicar indicadores como a evasão e reprovação escolar no ensino fundamental brasileiro. Eles então concluem que: "*Utilizar a mineração de dados educacionais possibilita a identificação prévia de aspectos que podem precisar de melhorias e investimentos mais adequados, aprimorando aspectos do ensino e mitigando problemas.*" (NASCIMENTO et al., 2018).

3 Materiais e Métodos

3.1 Dados Utilizados

Para realização deste trabalho, foi disponibilizado pela DAE os dados brutos referentes a inscrição dos requerentes de auxílio estudantil do ano de 2018. Por se tratar de dados sensíveis, nenhum tipo de informação de identificação dos solicitantes foi concedida. Para cada registo foi inserido um número de identificação para suprimir dados como CPF ou matrícula, mantendo assim a privacidade do requerente. Um colaborador da DAE detêm o mapeamento entre o número de identificação fictício e os dados reais, possibilitando a identificação do solicitante.

Os dados se referem a características socioeconômicas dos discentes e de seu grupo familiar, buscando assim, descrever quais são seus rendimentos, sua condição social e suas necessidades enquanto aluno da UNIFEI. Estes dados são obtidos pela DAE por meio de um formulário *online* preenchido pelos solicitantes à bolsa e foram fornecidos em formato de arquivo *xlsx*. Uma descrição mais detalhada sobre as características do conjunto de dados é apresentada na seção 4.2.

3.2 Metodologia

Existem diversos métodos para implementação de um projeto de ciência de dados. O estudo de Freire e Omar (2020) traz uma comparação entre 6 métodos, metodologias e frameworks na construção de sistemas computacionais analíticos-cognitivos, e seus resultados são baseados na adequação dos métodos em relação as estruturas básicas de um projeto. Os autores apresentam resultados mais expressivos para a metodologia CRISP-DM no desenvolvimento de projetos analíticos-cognitivos.

O trabalho de Sheth e Patel (2010) sugere que a melhor prática para construção de um projeto de EDM é a utilização da metodologia CRISP-DM, por conta de sua abrangência e generalização. Entretanto, RAMOS et al. (2020) propõe uma adaptação do modelo CRISP-DM para mineração de dados educacionais, o CRISP-EDM. A seguir, a metodologia CRISP-DM é apresentada de forma breve conforme a literatura, e as principais diferenças do modelo proposto por RAMOS et al. (2020) são apontadas na sequência.

A metodologia intitulada como Processo Padrão da Indústria Cruzada para Mineração de Dados (*Cross Industry Standard Process for Data Mining*, CRISP-DM) foi proposta por Chapman et al. (2000), e suas fases são ilustradas na Figura 4.

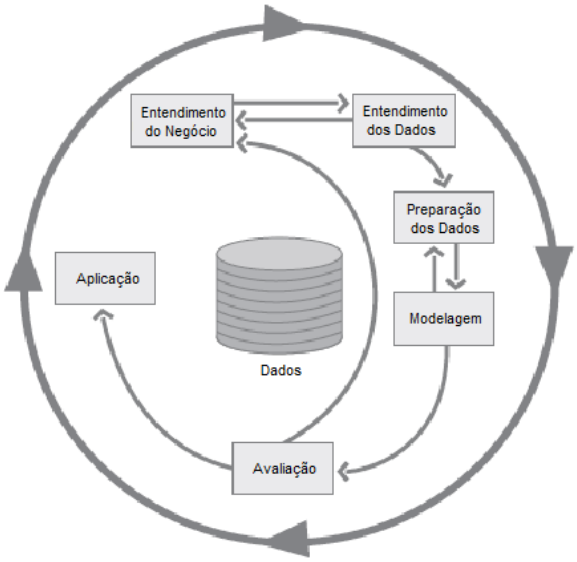


Figura 4: Fases do modelo de referência CRISP-DM, adaptado de Chapman et al. (2000).

O ciclo de vida do CRISP-DM é dividido em seis fases, sendo que a sequência de execução não é rígida, podendo ocorrer para frente ou para trás entre as diferentes fases no decorrer do processo (CHAPMAN et al., 2000). As fases são descritas a seguir:

1. **Entendimento do Negócio:** Nesta fase, deve-se identificar o problema do negócio que se deseja desenvolver. Possui três propósitos principais: (i) explicar a situação do ambiente e como o projeto pretende solucionar o problema; (ii) apresentar qual é o objetivo maior do projeto; (iii) evidenciar qual métrica que ditará o sucesso ou não do projeto.

2. **Entendimento dos Dados:** Neste momento deve-se arquitetar e estudar qual será a melhor forma de extrair informações dos dados, não sem antes haver revisões para identificar possíveis anomalias ou inconsistência no conjunto de dados, com o auxílio de tabelas e gráficos. Caso surjam dúvidas a aspectos técnicos dos dados ou algum impedimento, deve-se retornar a etapa anterior.

3. **Preparação dos Dados:** Esta fase visa preparar os dados para a modelagem. Serão selecionados os dados relevantes para o projeto, uma vez que nem todos os campos do conjunto de dados são pertinentes para alcançar o objetivo final. Haverá também o tratamento de tipos de variáveis incoerentes, para que se possa selecionar amostras aleatórias para treinos e testes.

4. **Modelagem:** Nesta fase realiza-se a construção do modelo, escolhendo as técnicas a serem utilizadas, em um primeiro momento executar em uma base exemplo e, caso necessário, voltar a etapa anterior para ajustes relevantes.

5. **Avaliação:** Junto ao grupo técnico de especialistas os resultados são analisados e é feito o levantamento de todas as possibilidades de variações que os dados podem apresentar,

de modo a testar os modelos. Caso os modelos não estejam adequados, deve-se retomar a primeira etapa para entender o negócio e os dados de forma mais clara.

6. **Aplicação:** A fase final é onde se faz uso de toda a análise desenvolvida no projeto. Essa análise pode ser desde a apresentação dos resultados da modelagem para a tomada de decisão até a aplicação do modelo em um outro conjunto de dados.

A metodologia CRISP-EDM proposta por RAMOS et al. (2020), embora muito similar a CRISP-DM, não corresponde aos propósitos desta pesquisa principalmente por contemplar questões acadêmicas relacionadas com o desempenho e não questões socioeconômicas. Portanto, a metodologia selecionada para estruturar o trabalho foi o modelo CRISP-DM.

3.3 Projeto de Experimentos

O objetivo dos experimentos é analisar a viabilidade do modelo de identificação de grupos de perfis de solicitantes de auxílio estudantil. Seguindo o CRISP-DM foram executadas as fases de 1 a 5. A execução da fase 6, referente a implantação do modelo proposto para o tomador de decisão, não faz parte do escopo deste trabalho.

A fase de Entendimento do Negócio foi desenvolvida por meio de pesquisas e reuniões com um colaborador da pesquisa que conhece o processo envolvido no negócio. Para fins de contextualização do problema abordado neste trabalho, o conteúdo relacionado a execução desta fase consta no Capítulo 1. Em seguida, na fase de Entendimento dos Dados, foi elaborado um dicionário de dados, identificadas as possibilidades de limpeza e realizada uma análise exploratória no conjunto de dados. Na fase de Preparação de Dados foi feita a limpeza do conjunto dados e a criação de novos atributos, a partir da combinação de atributos preexistentes. Um exemplo de novo atributo é o "Valor Total dos Bens Familiares", que é o somatório dos valores referentes aos bens do grupo familiar apresentados pelo solicitante. Em seguida, na fase de Modelagem, foram definidos os algoritmos de classificação não-supervisionados e supervisionados utilizados para a construção dos modelos de classificação. Por fim, na fase de Avaliação os modelos são analisados a partir de medidas de similaridade. Os métodos de aprendizado de máquina são comparados quanto a qualidade da classificação e desempenho. Os detalhes do que foi desenvolvido nas fases 1 a 5 da CRISP-DM constam no Capítulo 4.

Os experimentos foram executados em um notebook com processador Intel Core i7-4510U de 2.0 GHz, 12 GB de memória RAM, sistema operacional Windows 10 Home 64 bits. Os métodos não-supervisionados foram executados na linguagem Python versão 3.8.5 no ambiente de programação Jupyter Notebook. Os algoritmos de classificação não-supervisionada K-Means, Agglomerative e Birch foram executados de acordo com a implementação da biblioteca scikit-learn versão 0.23.2, o algoritmo K-Medoids da biblioteca scikit-learn-extra versão 0.2.0, e o algoritmo K-Prototypes da biblioteca Kmodes versão 0.11.0. Já para a tarefa de clas-

sificação supervisionada foi utilizado o *sotfware* de aprendizado de máquina WEKA versão 3.8.5, *sotfware* este de domínio público e implementado na linguagem de programação Java. A funcionalidade do WEKA utilizada foi a de classificação (*classify*), e o algoritmo utilizado foi o J48 que é uma implementação da Árvore de Decisão C4.5. Na Análise Exploratória e Pré-processamento foram utilizadas as bibliotecas Pandas versão 1.1.3, NumPy versão 1.19.2, Seaborn versão 0.11.1 e Matplotlib versão 3.4.2. Os parâmetros utilizados em cada método são descritos no Capítulo 5.

Ao final da execução, pretende-se obter diferentes agrupamentos de indivíduos utilizando diferentes métodos não-supervisionados para o agrupamento de dados (*clustering*). Estes métodos serão avaliados por índices de similaridade entre os grupos. Consequentemente, a tarefa de classificação será utilizada para identificar as regras implícitas para organização de cada grupo. Este conjunto de regras é o resultado da proposta deste trabalho que posteriormente deve ser avaliado pelo ponto de vista da ciência social, buscando validar ou não os resultados encontrados.

4 Desenvolvimento

Utilizando os materiais e métodos apresentados, as seções seguintes discorrem sobre cada etapa do método CRISP-DM. Estas etapas norteiam o presente trabalho e auxiliam no desenvolvimento da pesquisa, e na obtenção de um resultado autêntico.

4.1 Entendimento do Negócio

Em um período recente da história brasileira, as políticas de acesso às universidades foram ampliadas. Foram feitos investimentos para a construção de novas universidades, para o financiamento estudantil e assistência estudantil, além da implementação de políticas afirmativas. Este cenário evidenciou, dentro do ensino superior, problemas históricos intrínsecos ao país, como desigualdade social, racial e de gênero, impactando diretamente o número de alunos com algum tipo de vulnerabilidade acadêmica.

No entanto, algumas políticas de assistência estudantil, como auxílio permanência e alimentação, foram ampliadas na tentativa de mitigar uma possível evasão dos discentes das IFES. Segundo o PNAES cabe aos IFES definirem os critérios para concessão das bolsas. No âmbito da UNIFEI, tais políticas e recursos são gerenciados pela DAE e sua equipe técnica, composta por assistentes sociais, psicólogas e auxiliares, que seguem a norma definida em UNIFEI (2020), levando em consideração aspectos como moradia, renda e despesas do discente.

Atualmente, a medida que os recursos tornam-se cada vez mais escassos e o volume de solicitações aumenta, então formula-se a hipótese de que técnicas de EDM podem contribuir para o processo de tomada de decisão e distribuição de recursos por parte da DAE. Portanto, além do objetivo técnico, antes citado na introdução, de desenvolver um modelo de EDM que possa auxiliar os gestores acadêmicos, este projeto conta com o objetivo social de efetivamente buscar a permanência de grupos vulneráveis no ambiente acadêmico.

4.2 Entendimento dos Dados

Complementando a seção 3.1 sobre os Dados Utilizados, a base de dados considerada neste trabalho foi a do ano de 2018, que contém 961 registros. Seus atributos são listados na Tabela 2. Quanto a qualidade dos dados, verificou-se que havia muitos atributos sendo recebidos na entrada por um campo aberto, ou seja, seus valores são do tipo *string* e aberto a qualquer tipo de resposta. Este cenário ocasiona em maiores dificuldades para a preparação dos dados, descrita na seção 4.3, pois atributos importantes para análise tornam-se inconsistentes e por vezes até incoerentes, exigindo assim um esforço maior para o tratamento e limpeza destes

dados. Como exemplo, algumas respostas para o campo referente a renda bruta: (i) *"Em média R$ 980,00 , no entanto varias muitas vezes na média de R$ 200,00 para mais ou para menos.";* (ii) *"R$ 400,00 referente a bolsa de IC, vigente até setembro/outubro-2019.";* (iii) *" Valor total de proventos: +/- R$ 2.756,59Salario: R$ 1.822,73 Valor bruto de Janeiro a JaneiroObs: Tirando férias e 13º(os mesmos acontecem uma vez ao ano).".* Portanto nota-se que estes dados precisam de uma interpretação para a sua transformação.

A Tabela 2, referente ao dicionário de dados, descreve os atributos dos dados brutos, ou seja, sem a realização de tratamento. A coluna Atributo refere-se ao nome do atributo, a coluna Descrição faz uma breve descrição do atributo, a coluna Tipo refere-se ao tipo de dado e a coluna Limite de Valores descreve os possíveis valores do atributo. Os campos que estão apontados como "todo grupo familiar" referem-se aos atributos que se repetem de acordo com a quantidade de indivíduos do grupo familiar, incluindo o aluno, e os campos com "cada bem do grupo familiar" são repetidos para cada bem descrito pelo grupo familiar.

Tabela 2: Dicionário de Dados.

Atributo	**Descrição**	**Tipo**	**de Valores**
id_discente	Número identificador do discente.	Inteiro	Números Naturais
curso	Curso do discente.	String	Conjunto de cursos da UNIFEI
ano_ingresso	Ano de ingresso no curso atual.	Inteiro	Conjunto de anos até o ano de 2018
periodo_ingresso	Período de ingresso no curso.	Inteiro	[1 ; 2]
forma_ingresso	Forma de ingresso no curso.	String	Campo Aberto
campus	Em qual campus da universidade o discente cursa.	String	[Campus de Itabira ; Campus Sede Itajubá]
você exerce estágio remunerado?	Autoexplicativa	String	["Sim" ; "Não"]
alguém do seu grupo familiar cursa atualmente uma graduação?	Autoexplicativa	String	["Sim" ; "Não"]
você já possui uma graduação?	Autoexplicativa	String	["Sim" ; "Não"]
qual sua Procedência Escolar?	De qual das opções escolares o discente procede	String	["Em ESCOLA PÚBLICA (integralmente)" ; "FILANTRÓPICA" ; "PARTICULAR (com bolsa igual ou superior a 50% nos três anos do ensino médio)" ; "PARTICULAR (com bolsa inferior a 50%) ; PARTICULAR (sem bolsa)"]
qual a situação da Moradia do Aluno?	Autoexplicativa	String	["COM A FAMÍLIA /PARENTES /TERCEIROS" ; "DIVIDE ALUGUEL DE IMÓVEL COM AMIGOS/COLEGAS "; "SOZINHO EM IMÓVEL ALUGADO/FINANCIADO" ; "SOZINHO EM IMÓVEL PRÓPRIO (QUITADO)"]
valor mensal (para opções 'Alugada', 'Financiada')	Caso a opção anterior tenha sido Alugada ou Financiada, quanto é gasto mensalmente	String	Campo Aberto

Continua na página seguinte.

Tabela 2 – Continuação da página anterior

Atributo	Descrição	Tipo	Limite de Valores
se divide este valor, com quantas pessoas?	Referente ao valor do campo anterior	String	[“1” ; “2” ; “3” ; “4” ; “5 ou mais”]
qual a situação da Moradia do Grupo Familiar?	A mesma situação porém para o grupo familiar	String	[“ALUGADA” ; “CEDIDA OU HERANÇA” ; “PRÓPRIA E QUITADA” ; “PRÓPRIA EM PAGAMENTO (FINANCIADA)”]
valor mensal (para opções ’Alugada’,’Financiada’)	Referente ao campo anterior	String	Campo Aberto
qual o principal meio de transporte que você utiliza para vir até a Universidade? (campo aglutinado)	Autoexplicativa	String	[“Transporte coletivo” ; “Carona” ; “A pé/bicicleta” ; “Carro/moto (próprio) ” ; “Transporte locado ”]
qual o valor mensal em gasto com transporte?	Autoexplicativa	String	Campo Aberto
qual a distância da sua residência até a universidade (em Km)?	Autoexplicativa	String	[“até 2,0 km” ; “entre 2,0 e 5,0 km” ; “mais de 5,0 km ”]
qual o Estado Civil? (todo o grupo familiar)	Autoexplicativa	String	[“Solteiro” ; “Casado” ;“União estável” ; “Separado/divorciado”]
qual a Ocupação/Profissão? (todo grupo familiar)	Autoexplicativa	String	Campo Aberto
qual a renda bruta mensal? (todo grupo familiar)	Autoexplicativa	String	Campo Aberto
se possui doença crônica, indique qual: (todo grupo familiar)	Autoexplicativa	String	Campo Aberto
idade: (todo grupo familiar)	Autoexplicativa	String	Campo Aberto
parentesco: (todo grupo familiar)	Autoexplicativa	String	[“Filho(a)” ; “Mãe” ;“Pai” ;“Avô(ó)” ;“Irmão(ã)” ;“Tio(a)” ;“Outros”]
escolaridade: (todo grupo familiar)	Autoexplicativa	String	[“Fundamental Completo” ; “Fundamental Incompleto” ; “Médio Completo” ; “Médio Incompleto” ; “Superior Completo” ; “Superior Incompleto” ; “Pós-graduação”]
outros rendimentos recebidos pelo grupo familiar: (campo aglutinado)	Autoexplicativa	String	[“Mesada” ; “Aluguel, arrendamento” ; “Auxílio de parentes ou amigos” ; “Outros”]
valor mensal: (de outros rendimentos do grupo familiar)	Referente ao campo anterior	String	Campo Aberto
possui veículos?	Autoexplicativa	String	[“Sim” ; “Não”]
se sim, quantos?	Autoexplicativa	String	Campo Aberto
valor total do(s) IPVA(s) sem desconto:	Autoexplicativa	String	Campo Aberto
descrição: (cada bem do grupo familiar)	Descrição dos bens patrimoniais do grupo familiar	String	Campo Aberto
município: (cada bem do grupo familiar)	Município em que o bem está registrado	String	Campo Aberto
valor de mercado: (cada bem do grupo familiar)	Valor de mercado do campo anterior	String	Campo Aberto

4.3 Preparação dos Dados

A aplicação de regras para limpeza dos dados é muito importante no processo de MD, principalmente quando os dados tem origem manual, isto é, humana. A etapa de definição e execução das regras visa garantir que o conjunto de dados obtido após a limpeza corresponda de forma autêntica aos dados brutos. Neste trabalho, esta etapa foi executa manualmente de acordo com as regras definidas na Tabela 3. Isso se deu em função da complexidade envolvida no tratamento dos dados, conforme descrito na seção 4.2.

As regras foram definidas a partir de aspectos técnicos e conhecimentos tácitos. Os aspectos técnicos são definidos a partir das premissas necessárias para o desenvolvimento de um processo de MD, isto significa, alteração dos tipos dos dados, remoção de incoerências, ajuste de inconsistências, discretização, entre outros. Já os aspectos relacionados ao conhecimento tácito, mais especificamente sobre o negócio, contaram com a assessoria de um colaborador da instituição com experiência em assistência estudantil e administração acadêmica. O nome e descrição das regras definidas para limpeza dos dados são:

- **Encurtamento da *String*:** Foi realizado o processo de tornar a *String* menor, sem perda de informação. Isto foi feito para que, em uma posterior análise, esses atributos tenham uma visualização mais coesa e agradável. Pro exemplo: “Em ESCOLA PÚBLICA (integralmente)” $\rightarrow$ “Pública”.

- **Alteração de tipo (*String->int*):** Dados que estavam em formato de *String*, porém possuiam características numéricas importantes, foram interpretados e transformados em tipo numérico inteiro. Por exemplo: “Em torno de R$500” $\rightarrow$ “500”.

- **Aglutinação de atributos:** Atributos de múltipla escolha que poderiam estar em apenas um atributo, sem perda de qualquer informação, foram agrupados. Nestes casos, sequência de perguntas de sim ou não, onde apenas uma das opções poderia ser verdadeira, foi aglutinado em um atributo correspondendo a resposta positiva. Por exemplo o atributo: “Qual o principal meio de transporte que você utiliza para vir até a Universidade? ”; com as respostas: “Transporte coletivo”, “A pé/bicicleta ”, “Carro/moto (próprio) ”, foram aglutinados apenas em suas respostas.

- **Discretizado por doenças:** Esta transformação se refere especificamente as doenças crônicas apontadas pelo aluno ou pelo grupo familiar. Para a discretização, foram definidos três categorias de doenças: **grupo 0** - pessoas sem doenças crônicas ou doenças leves, exemplo: rinite; **1** - pessoas com doenças crônicas não incapacitantes, exemplo: diabetes; **2** - pessoas com doenças crônicas incapacitantes, exemplos: paralisia infantil. As categorias usadas na discretização e as respectivas doenças foram definidas pelos pesquisadores.

Tabela 3: Alterações nos dados.

Atributo	**Alteração**
Qual sua Procedência Escolar?	Encurtamento da *String*
Qual a situação da Moradia do Aluno?	Encurtamento da *String*
Valor mensal (para opções 'Alugada', 'Financiada)(do aluno)	Alteração de tipo (*String->int*)
Qual a situação da Moradia do Grupo Familiar?	Encurtamento da *String*
Valor mensal (para opções 'Alugada', 'Financiada) (do grupo familiar)	Alteração de tipo (*String->int*)
Qual o principal meio de transporte que você utiliza para vir até a Universidade?	Aglutinação de atributos
Qual o valor mensal em gasto com transporte?	Alteração de tipo (*String->int*)
Qual a distância da sua residência até a universidade (em Km)?	Encurtamento da *String*
Idade (para o próprio aluno e todos os indivíduos do grupo familiar)	Alteração de tipo (*String->int*)
Qual a renda bruta mensal (o próprio aluno)?	Alteração de tipo (*String->int*)
Se possui doença crônica, indique qual: (para o próprio aluno e todos os indivíduos do grupo familiar)	Discretizado por doenças
Renda bruta mensal (para todos os indivíduos do grupo familiar)	Alteração de tipo (*String->int*)
Outros rendimentos recebidos pelo grupo familiar:	Aglutinação de atributos
Valor mensal (outros rendimentos):	Alteração de tipo (*String->int*)
Valor total do(s) IPVA(s) sem desconto:	Alteração de tipo (*String->int*)
Valor de mercado: (para cada bem do grupo familiar)	Alteração de tipo (*String->int*)

Após executar a limpeza no conjunto de dados, novos atributos foram gerados por meio da combinação de outros atributos da base de dados. A criação destes campos foi motivada pela redução da dimensão do conjunto através da aglutinação dos atributos, sendo mais apropriado para a execução dos métodos de MD (DASH et al., 1997). O conjunto de dados foi reduzido de 113 para 22 atributos, aproximadamente 82% de redução. A Tabela 4 contém a relação dos novos atributos e a sua constituição.

Tabela 4: Novos Atributos.

Atributo	**Origem**
Quantos filhos o solicitante possui?	Somatório de quantos filhos são apresentados no grupo familiar no atributo parentesco
Renda per capita	Somatório da Renda bruta mensal de cada indivíduo do grupo familiar dividido pela quantidade de indivíduos no grupo familiar
Despesas per capita	Somatório de todas as despesas mensais apresentadas pelo grupo familiar dividido pela quantidade de indivíduos no grupo familiar
Familiares com Superior Completo ou Pós	Contagem de indivíduos no grupo familiar com ensino superior ou pós-graduação
Quantidade de indivíduos com doença grave no grupo familiar	Contagem de indivíduos no grupo familiar com doenças incapacitantes (grupo 2)
Valor Total dos bens familiares	Somatório do valor de mercado informado pelo discente para cada bem do grupo familiar

Ao fim da execução das ações de preparação dos dados, é obtido o conjunto de dados reduzido, cujos atributos são listados a seguir: (1) “id_discente”; (2) “curso” ; (3) “ano_ingresso” ; (4) “periodo_ingresso” ; (5) “forma_ingresso” ; (6) “campus” ; (7) “Você exerce estágio remunerado?” ; (8) “Alguém do seu grupo familiar cursa atualmente uma graduação?” ; (9) “Você já possui uma graduação?” ; (10) “Qual sua Procedência Escolar?” ; (11) “Qual a situação da Moradia do Aluno?” ; (12) “Qual a situação da Moradia do Grupo Familiar?” ; (13) “Qual o principal meio de transporte que você utiliza para vir até a Universidade?” ; (14) “Qual a distância da sua residência até a universidade (em Km)?” ; (15) “Qual o Estado Civil (o próprio aluno)?” ; (16) “Qual a Ocupação/Profissão (o próprio aluno)?” ; (17) “Quantos filhos o solicitante possui?” ; (18) “Quantidade de individuos com doença grave no grupo familiar” ; (19) “Renda per capita” ; (20) “Despesas per capita” ; (21) “Familiares com Superior Completo ou Pós” ; (22) “Valor Total dos bens familiares”.

4.3.1 Seleção dos Atributos

A seleção dos atributos é uma tarefa importante em um processo de mineração de dados, logo que a medida que a dimensionalidade do conjunto de dados aumenta, cresce também as incertezas sobre seus resultados. Isso significa que um conjunto de dados com muitos atributos tende a gerar um resultado impreciso ou pouco informativo. Portanto, as etapas de aglutinação de atributos, geração de novos atributos que representem um conjunto de campos, análise exploratória, entre outros processos, são essenciais para o resultado final do processo.

Neste trabalho, a escolha foi conduzida com base na NORMA 2.3.03 (UNIFEI, 2020)

instituída pelo Conselho de Ensino, Pesquisa, Extensão e Administração (CEPEAD) da UNIFEI. O Apêndice A da norma define quais os parâmetros e as pontuações que devem auxiliar a equipe da DAE a ranquear os requerentes. Estas pontuações foram importantes para definir os atributos selecionados para execução do processo de mineração.

A seleção dos atributos foi dividida em dois grupo, um grupo que reúne apenas atributos numérico (seis no total), denominado **6Num**, e outro que reúne os seis atributos numéricos e três categóricos (textuais), denominado **6Num3Cat**. Esta divisão foi adotada para atender os requisitos de entrada dos métodos de agrupamento, considerando que alguns deles operam apenas com variáveis numéricas, como o K-Means, e outros com variáveis numéricas e categóricas, como o K-Prototypes. A seguir a Tabela 5 apresenta os atributos selecionados para cada grupo.

Tabela 5: Atributos selecionados para o grupo 6Num e 6Num3Cat.

Atributo	**6Num**	**6Num3Cat**
Renda per capita	x	x
Despesas per capita	x	x
Valor Total dos bens familiares	x	x
Familiares com Superior Completo ou Pós	x	x
Quantidade de individuos com doença grave no grupo familiar	x	x
Quantos filhos o solicitante possui?	x	x
Qual sua procedência escolar?		x
Qual a situação da moradia do aluno?		x
Qual a situação da moradia do grupo familiar?		x

Considerando o conjunto de dados selecionado foi realizada uma Análise Exploratória para um melhor entendimento das características dos requerentes de assistência estudantil. Esta etapa esta descrita no Capítulo 5.

4.3.2 Pré-Processamento

O pré-processamento é uma tarefa fundamental no processo de mineração de dados, já que, os resultados encontrados ao final desta etapa correspondem com o conjunto de dados de entrada para os métodos de aprendizado. Isso significa, por exemplo, que atributos com diferenças na ordem de grandeza ou atributos não categorizados podem sujeitar alguns métodos a resultados inconsistentes, uma vez que são usadas medidas de distância para inferir a similaridade entre as entidades. Esta etapa prevê também o uso de técnicas para atenuar os *outliers*, logo que seus valores absolutos podem enviesar um determinado atributo.

Neste trabalho serão utilizados dois tipos de métodos de agrupamento, alguns que lidam apenas com atributos numéricos e outros admitem dados numéricos e categóricos. O pré-processamento será realizado apenas nos atributos numéricos, tendo em vista que os atributos não numéricos estão de certo modo normalizados por já estarem categorizados em decorrência da execução da regra de encurtamento de *string* apresentado na Tabela 3. Sendo assim, serão apresentados abaixo os quatro diferentes pré-processamentos realizados nesse trabalho:

4.3.2.1 Normalização MIN-MAX

A normalização MIN-MAX é uma técnica para transformar os valores, levando em consideração o valor máximo e mínimo dos registros de um determinado atributo. Porém, como se pode observar nas Figuras 12 e 13, os atributos numéricos contínuos, como renda e despesas, apresentam um número considerável de *outliers*, sendo necessário um procedimento para normalização dos atributos.

Neste procedimento, inicialmente é feita a multiplicação dos valores dos atributos numéricos contínuos por *ln* (logaritmo natural). Isto se fez necessário por conta das variáveis abrangerem valores de diversas ordens de magnitude, ou seja, suas distribuições são similares a uma "lei da potência", o que significa que a grande maioria dos valores são pequenos e poucos são muito grandes. Este tipo de distribuição é estudada em escala logarítmica, logo que transforma uma grande diferença em uma escala $10^5 - 10^2$, em uma menor escala como 5 - 2, tornando os valores comparáveis.

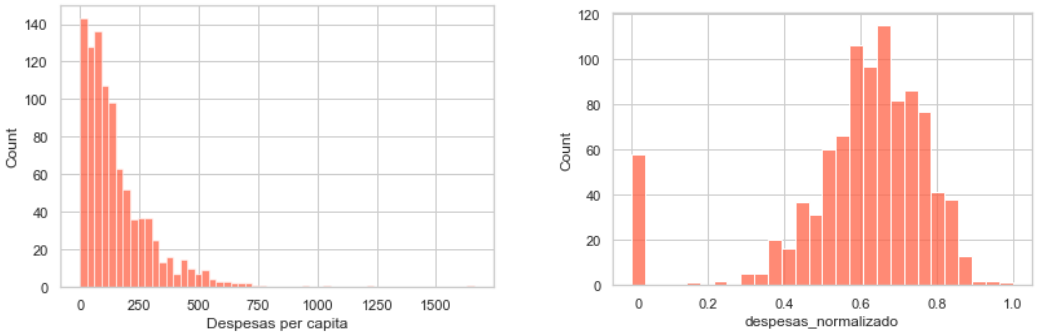
Após este procedimento foi executada a normalização MIN-MAX, cuja fórmula está descrita na Equação 4.1.

$$x' = \frac{x - min(x)}{max(x) - min(x)} \tag{4.1}$$

onde *x* representa o valor precedente; *min(x)* o menor valor do atributo entre todos os registros; *max(x)* o maior valor do atributo entre todos os registros; *x'* o valor normalizado.

A normalização foi realizada em todos os atributos numéricos, ou seja, o subconjunto que compreende a seleção 6Num ta Tabela 5. Na Figura 5, nota-se o resultado da normalização

quando posicionamos os gráficos lado a lado. Na Figura 5a observa-se o histograma dos dados sem a etapa de pré-processamento, o qual apresenta um comportamento tipo declive a direita, e na Figura 5b o resultado após a etapa descrita nesta subseção. Observa-se que a curva ilustrada no gráfico normalizado é similar a do gráfico da Distribuição Gaussiana (Normal) e que seus valores variam de 0 a 1.



(a) (b)

Figura 5: Gráficos para comparação dos dados brutos (*raw data*) e dados pré-processados (*pre-processed data*) do atributo “Despesas per capita”.

4.3.2.2 Normalização MIN-MAX com pesos

O segundo pré-processamento avaliado foi a normalização MIN-MAX, no entanto ao final do processo foi multiplicada uma constante em alguns atributos para destacar sua importância em relação aos demais atributos a partir de pesos. Este procedimento é similar ao apresentado na subseção anterior, diferindo apenas por considerar pesos.

Os atributos escolhidos como mais relevantes conforme os critérios descritos na seção 4.3.1, foram os atributos: “Despesas per capita” e “Renda per capita”. Estes atributos ao final do processo de normalização MIN-MAX foram multiplicados pela constante 2, na tentativa de reproduzir um peso duas vezes maior com relação aos outros atributos.

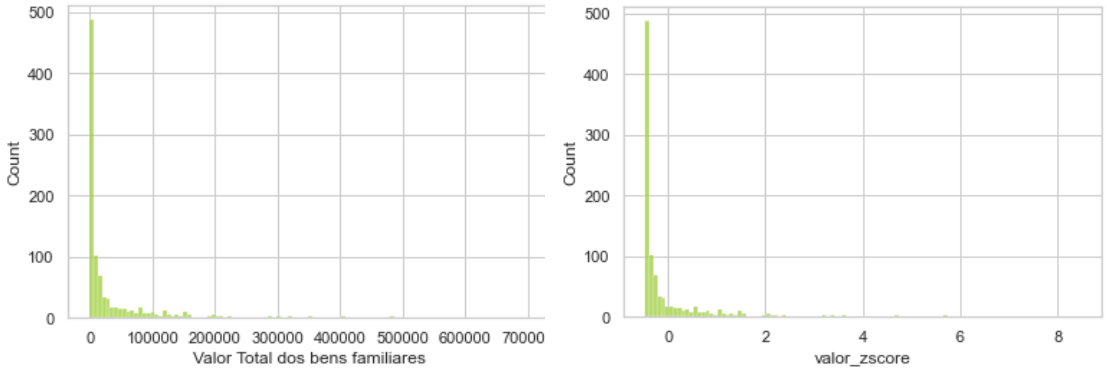
4.3.2.3 z-score

Diferente das estratégias anteriores, o z-score ou escore padrão procura padronizar o conjunto de dados baseando-se no desvio padrão. Essa técnica ajusta aproximadamente 97% de dados entre os valores -1,96 e 1,96, valores que extrapolem estes limites são considerados incomuns. A Equação 4.2 ilustra a fórmula do z-score.

$$Z = \frac{x - \mu}{\sigma} \tag{4.2}$$

onde x representa o valor precedente; μ a média dos valores do atributo entre todos os registros; σ o desvio padrão dos valores do atributo entre todos os registros; Z o valor normalizado.

O z-score também é um método de normalização, que se baseia no desvio padrão. A aplicação da fórmula z-score distingue-se, entre outras coisas, do método de normalização pois não altera a distribuição do conjunto de dados, e sim sua escala. Na Figura 6a temos o histograma referente aos *raw datas* do atributo 'Valor Total do bens familiares' e na Figura 6b o histograma após a execução do z-score.



(a) (b)

Figura 6: Gráficos para comparação *raw data* e *pre-processed data* para o atributo Valor total dos bens familiares com pré-processamento z-score.

4.4 Modelagem

O modelo definido neste trabalho busca a identificação de grupos de solicitantes de auxilio estudantil com perfis similares, além da classificação de um novo solicitante em um desses grupos. A estrutura que fundamenta o modelo é baseada na metodologia do ciclo de descoberta de conhecimento em dados, descrito principalmente por Fayyad et al. (1996), Han et al. (2011) e Cortes et al. (2002). Sua abordagem é do tipo *top-down*, logo que seu objetivo é validar uma hipótese, previamente apresentada neste trabalho.

A Figura 7 ilustra a iteração entre as etapas do modelo e os métodos utilizados nas fases de Pré-Processamento, Aprendizado Não-supervisionado (*Unsupervised Learning*) e de Aprendizado Supervisionado (*Supervised Learning*).

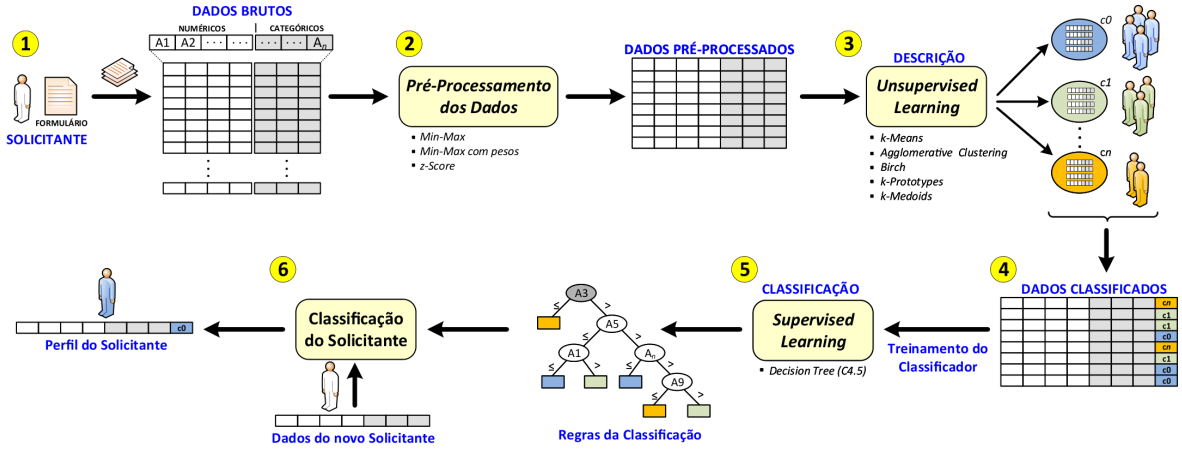


Figura 7: Modelo do processo de mineração utilizado.

A primeira etapa compreende o recebimento via relatório, dos dados socioeconômicos referentes aos solicitantes, seguido pela limpeza e seleção dos atributos. Posteriormente na segunda etapa, é realizado o pré-processamento deste conjunto de dados, resultando em um novo conjunto pré-processado para cada tipo de abordagem utilizada, são elas: MIN-MAX, MIN-MAX com pesos e z-score. Estas duas etapas são indispensáveis para um resultado coerente ao final do processo de mineração de dados, estando descritas na seção 4.3.2.

Com a definição conjunto de dados para mineração, a terceira etapa consiste na execução dos algoritmos de agrupamento, sendo que os algoritmos exclusivamente numéricos como K-Means, Agglomerative Clustering e Birch utilizam o conjunto de dados 6Num apresentado na Tabela 5, e os algoritmos que lidam com dados numéricos e categóricos, K-Medoids (com Gower) e K-Prototypes, fazem uso do conjunto 6Num3Cat também presente na Tabela 5.

Na etapa 4, no registro de cada solicitante é adicionado a classe (grupo) que ele pertence. Estes dados serão utilizados etapa 5 para o treinamento do algoritmo de classificação supervisionada, para o caso deste trabalho a Árvore de Decisão J48, que é uma implementação do algoritmo C4.5. O algoritmo cria uma árvore de decisão, que descreve as regras para inclusão

de um indivíduo em um determinado grupo. E por fim, a etapa 6 que classifica novos indivíduos nos grupos existentes de acordo com as regras do classificador.

4.5 Avaliação

A etapa de avaliação visa descrever como os métodos de classificação não-supervisionada e supervisionada foram executados, a definição dos parâmetros, e sobre que índices os resultados foram comparados.

Na tarefa de aprendizado não-supervisionado foram executados algoritmos de agrupamento, cuja característica é organizar os dados em grupos, a partir de um determinado índice. Foram selecionados os algoritmos K-Means, Agglomerative Clustering e Birch, que trabalham apenas com valores numéricos, e os métodos K-Prototypes e K-Medoids (com a medida de distância Gower) que trabalharam com dados numéricos e categóricos. Os algoritmos foram escolhidos de modo a considerar nos experimentos diferentes abordagens de agrupamento, como os Particionais e Hierárquico.

Para cada algoritmo de agrupamento são exigidos parâmetros diferentes para sua execução. Esta situação gera uma necessidade de se avaliar o resultado dos agrupamentos a partir do valor de índices de similaridade, para que os parâmetros escolhidos para cada algoritmo sejam os melhores possíveis. Estes índices são importantes pois os métodos não-supervisionados são difíceis de avaliar por conta da não existência de um modelo de metas para comparar aos resultados.

Para os algoritmos que trabalham exclusivamente com dados numéricos, foram escolhidos 3 índices de validação de agrupamento (clustering evaluation). Suas fórmulas são descritas e explicadas detalhadamente no trabalho de Desgraupes (2017). Abaixo segue uma breve explicação sobre os índices:

- **Silhouette:** é uma medida da similaridade média de um objeto dentro de um grupo (*intra-cluster*) em relação ao grupo mais próximo em que aquele objeto não faz parte. Seus valores variam de -1 a 1, sendo que quanto mais próximo do valor 1 cada objeto está mais próximo de seu grupo do que do grupo vizinho mais perto, e -1 caso contrário.
- **Davies-Bouldin:** é baseado em uma proporção entre as distâncias entre os pontos dentro de um grupo e as distâncias entre os centros de cada grupos. Seus valores podem ir de 0 a qualquer valor maior que 0, onde o quão mais próximo de 0 representa uma melhor avaliação segundo o índice.
- **Dunn:** seu objetivo principal é em avaliar se os diferentes grupos encontrados são compactos e apresentam pouca variabilidade entre os pontos. Seu valores variam de 0 a 1, onde quanto mais próximo do valor 1 os grupos encontrados são melhor avaliados.

Já para os métodos que executaram dados numéricos e categóricos, não foram encontrados na literatura índices bem estabelecidos para avaliação interna ou externa dos *clusters* obtidos. Portanto, foi utilizado a soma das distâncias quadradas, para determinar o melhor número de grupos para os métodos K-Prototypes e K-Medoids. A soma dos erros quadrados é uma medida estatística de discrepância entre os dados. A implementação utilizada destes algoritmos possibilita obter o valor desta medida nos resultados.

Para a etapa de classificação supervisionada foi utilizado o *software* WEKA, mais especificamente a funcionalidade de classificação (*classify*). O classificador utilizado foi a Árvore de Decisão J48, com o número mínimo de objetos igual a 10, o qual define o menor valor possível de indivíduos em um nó folha da árvore. Para realização dos testes foi escolhida a técnica *Cross Validation* (HASTIE et al., 2009), na qual o conjunto de dados é particionado em k subconjuntos de tamanhos similares, e em seguida, realiza uma bateria de treinamentos e testes aleatórios nestes subconjuntos. O valor k escolhido foi 10.

Estes parâmetros foram utilizados para todos os resultados obtidos no processo de agrupamento. Esta decisão é em virtude dos bons resultados apresentados nesta etapa, e por conta da maior relevância da etapa de agrupamento, uma vez que o objetivo mais significativo da pesquisa é a identificação de grupos de indivíduos com características semelhantes. O modelo proposto neste trabalho possibilita a utilização de outros métodos de classificação desde que sejam respeitados os formatos de entrada e saída de dados.

Para a avaliação do classificador, foram utilizados índices que são descritos de forma resumida a seguir. Uma descrição mais detalhada sobre cada um pode ser encontrada em Sokolova e Lapalme (2009) e Cohen (1960).

- **Kappa:** é uma estatística usada para medir a confiabilidade do resultado encontrado pelo classificador. Leva em consideração a possibilidade da previsão ocorrer por acaso, sendo que valores acima de 0.8 são considerados ótimos.

- **Acurácia Balanceada:** é um índice calculado com base na matriz de confusão que referenciam os erros e acertos cometidos pelo classificador. Este índice faz um cálculo em cima dos verdadeiros positivos e verdadeiros negativos encontrados na matriz, resultando em um valor não influenciado pelo desbalanceamento das classes.

- **Média AUC:** o termo AUC significa *Area Under the Curve*, e consiste do cálculo da área abaixo da curva ROC (*Receiver Operating Characteristic*) que é uma representação gráfica que ilustra o desempenho de um classificador. Em um sistema multiclasses as curvas são geradas para cada classe, portanto são calculados os AUCs para cada classe e em seguida é determinado a média de todo o conjunto.

- **Tamanho da Árvore:** este não é um índice de qualidade mas sim um valor que representa o tamanho da árvore de decisão gerada pelo classificador. Auxilia na identificação da

complexidade das regras que definem um grupo.

Com base nas etapas descritas no Desenvolvimento e nos índices apresentados, os resultados da pesquisa são ilustrados, descritos e discutidos no capítulo seguinte.

5 Resultados

Este capítulo discorre sobre os resultados obtidos na pesquisa. Primeiramente é apresentada a análise exploratória que consiste em explorar o conjunto de dados de forma gráfica e descritiva para entender suas características. Em seguida, é apresentado o resultado dos método de agrupamento e de classificação. Estes geram diferentes resultados que são ilustrados nas seções 5.2 e 5.3, que têm a finalidade de apresentar e comparar seus resultados e apontar os mais relevantes quanto ao objetivo da pesquisa, por meio dos índices e pela percepção do pesquisador.

5.1 Análise Exploratória

A etapa de Análise Exploratória consiste em examinar estatisticamente os dados, organizá-los e sintetizá-los de modo a obter informações gráficas simples e claras que ilustrem de algum modo um conhecimento sobre os dados.

A primeira tarefa foi visualizar a matriz de correlação (Figura 8) que mostra o coeficiente de correlação entre os dados numéricos selecionados para execução. Este coeficiente varia de -1 a 1, sendo que um coeficiente positivo e próximo de 1 representa uma correlação direta ou em uma mesma direção, e um coeficiente negativo e próximo de -1 representa uma correlação inversa ou em direção oposta. Os dois atributos não apresentam nenhuma correlação quanto mais próximo ao valor 0.

	Quantidade de individuos com doença incapacitante no grupo familiar	Quantos filhos o solicitante possui?	Renda per capita	Despesas per capita	Familiares com Superior Completo ou Pós	Valor Total dos bens familiares
Quantidade de individuos com doença incapacitante no grupo familiar	1.000000	-0.039856	-0.028979	-0.051690	0.003605	0.031329
Quantos filhos o solicitante possui?	-0.039856	1.000000	-0.074499	0.047585	-0.008589	-0.038683
Renda per capita	-0.028979	-0.074499	1.000000	0.151156	0.201378	0.109098
Despesas per capita	-0.051690	0.047585	0.151156	1.000000	0.073596	-0.064771
Familiares com Superior Completo ou Pós	0.003605	-0.008589	0.201378	0.073596	1.000000	0.084443
Valor Total dos bens familiares	0.031329	-0.038683	0.109098	-0.064771	0.084443	1.000000

Figura 8: Matriz de Correlação

De forma geral, os valores obtidos na tabela não apresentaram altos índices de correlação. Nota-se também que os maiores valores de correlação foram entre a “Renda per capita” e a “Despesas per capita” com valor 0.151; a “Renda per capita” com “Familiares com Superior Completo ou Pós” com valor 0.201; e a “Renda per capita” e o “Valor Total dos bens familiares” com valor 0.109. Isso indica que a capacitação do núcleo familiar tem uma relação direta com com a renda do grupo. Esta correlação também é observada na Figura 9, que consiste em um gráfico *boxplot* onde o eixo *x* representa os grupos do atributo “Familiares com Superior Completo ou Pós”, e o eixo *y* o atributo “Renda per capita”. Cada diagrama divide o grupo em quatro

partes iguais com 25% dos registros do grupo, e suas médias são grafadas no gráfico. É possível observar que a medida que o número de indivíduos com superior completo no grupo familiar aumenta, a média da renda do grupo familiar também cresce, com exceção do grupo com 3 indivíduos com superior completo, que pode ser atributo a pequena quantidade da amostra, de apenas 2 registros.

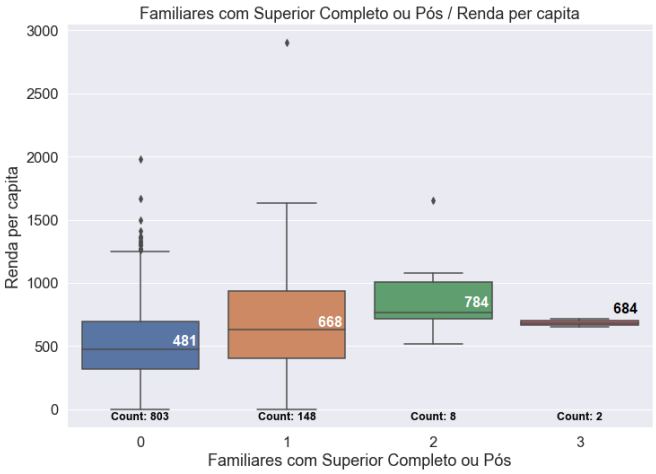


Figura 9: Gráfico *boxplot* para os grupos "Familiares com Superior Completo ou Pós" em relação a "Renda per capita".

A correlação ilustrada no gráfico da Figura 9 pode ser lógica para o entendimento de um ser humano logo que indivíduos graduados tendem a ter uma renda maior do que indivíduos com escolaridade inferior. Contudo, a análise exploratória contribui tanto para reafirmar as concepções do senso comum quanto para mostrar o contrário. Como por exemplo o gráfico apresentado na Figura 10, que consiste na correlação do atributo 'Qual sua Procedência Escolar?' com a média da "Renda per capita" e a média do "Valor Total dos Bens Familiares". Mediante ao encurtamento de *string* realizado na etapa de limpeza, o valor "PARTICULAR <50%" se refere a "PARTICULAR (com bolsa inferior a 50%)" e o valor "PARTICULAR >50%" se refere a "PARTICULAR (com bolsa igual ou superior a 50% nos três anos do ensino médio)".

O senso comum diria que estudantes oriundos do ensino público tendem a ter a renda familiar inferior a estudantes vindos de escolas particulares, porém o gráfico da Figura 10 apresenta um equilíbrio entre essas duas classes. Logo que entre todas as opções de procedência o valor médio de renda varia entre $\approx$ R$ 550,00 e $\approx$ R$ 750,00, apresentando seus menores valores para ensino pública e ensino particular sem bolsa. Entretanto é possível notar um desequilíbrio quanto a média do patrimônio familiar. Estes fatores podem promover a discussão sobre a validade do uso da procedência escolar como fator decisório. Este gráfico por sua vez não pode ser generalizado, já que a base de dados utilizada neste trabalho é um recorte específico de estudantes da Universidade Federal de Itajubá, requerentes de auxílio estudantil no ano de 2018.

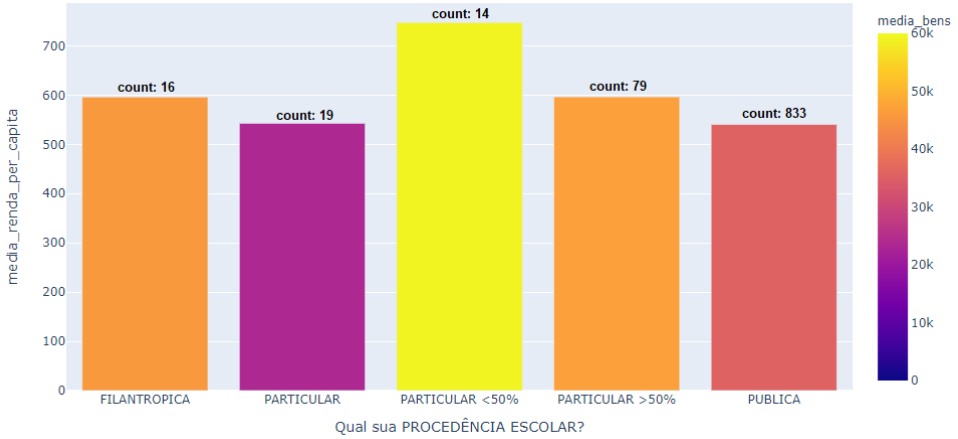


Figura 10: Gráfico de Correlação 2

Outro fator que pode ter influenciado este resultado é a grande diferença percentual em relação a procedência escolar entre os requerentes, ilustrado no gráfico da Figura 11. Observa-se que 86.7% dos solicitantes são oriundos do ensino público, e a segunda maior porcentagem, de 8.2%, se refere aos estudantes oriundos do ensino particular, porém com bolsa maior que 50% do valor.

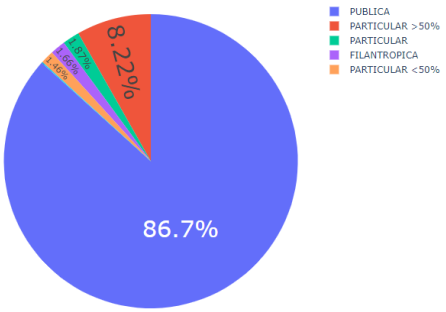
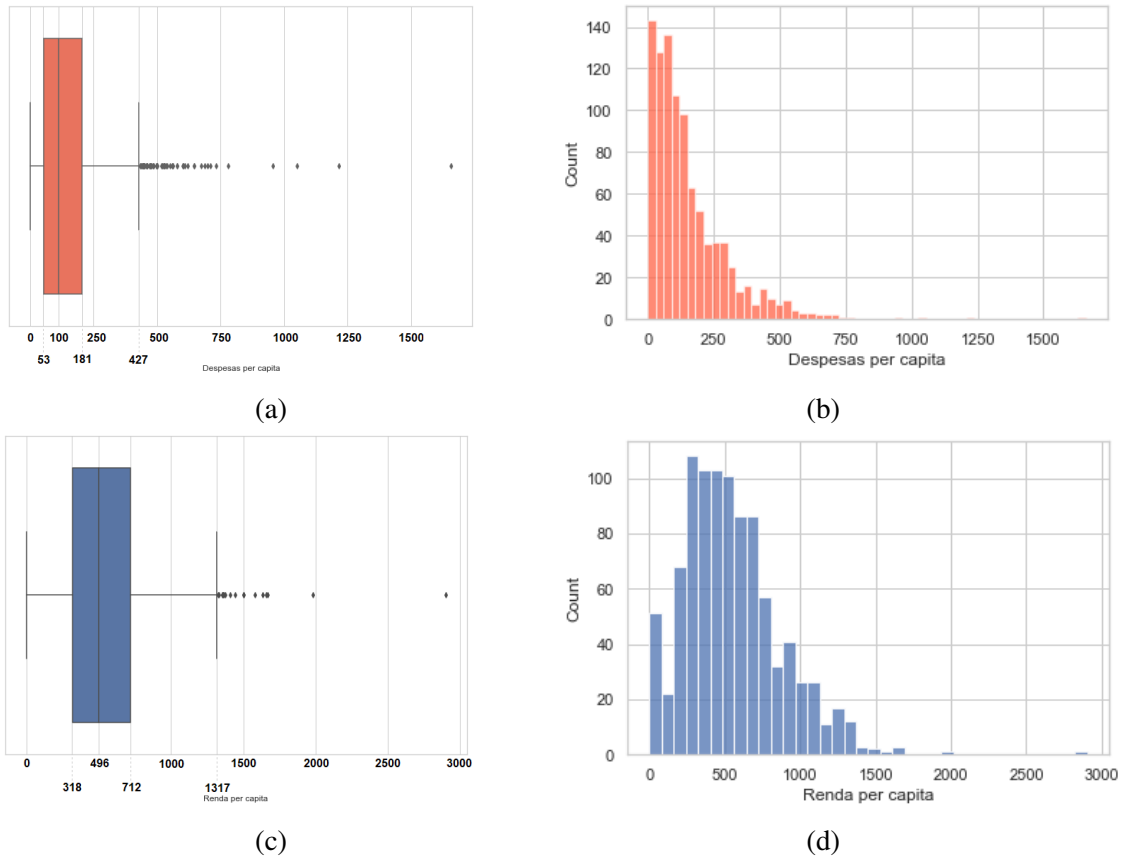


Figura 11: Gráfico da porcentagem da procedência escolar.

Os gráficos da Figura 12 ilustram a distribuição dos valores para os atributos considerados mais relevantes do conjunto de dados. O histograma visa ilustrar a distribuição de frequência dos valores dos atributos e a quantidade de registros que apresentam aquele valor. Já o *boxplot* apresenta como estão divididos cada quartil do conjunto de dados. Os pontos sobressalentes são considerados *outliers*, que são pontos que divergem da média dos valores dos quartis.

O atributo Despesa per capita, Figuras 12a e 12b, no histograma apresenta uma distribuição assimétrica a direta e no *boxplot* conta com 59 *outliers* com uma média de $\approx$ R$ 577,00, discrepantes da média geral deste atributo que é de $\approx$ R$ 110,00. Já o atributo Renda per capita, Figura 12c e 12d, apresenta uma distribuição próxima a normal com média geral de $\approx$ R$ 500,00, contando com 17 *outliers* com média de $\approx$ R$ 1573,00.

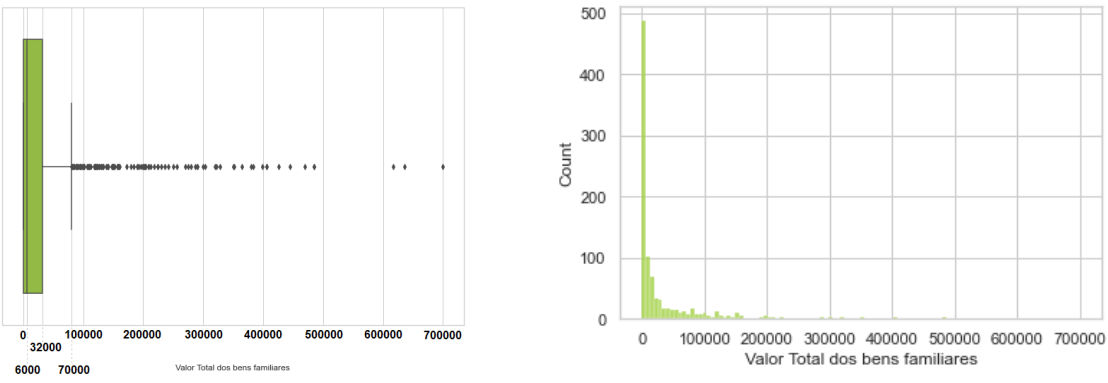


(a) (b)

(c) (d)

Figura 12: Gráficos *boxplot* e histograma do atributo Despesas per capita e Renda per capita.

Todavia o atributo que referencia o valor do patrimônio familiar, presente na Figura 13, exibe uma distribuição desequilibrada dos valores, onde 87% das ocorrências apresentam uma média de aproximadamente ≈ R$ 12.000,00 e os 13% de *outliers* com média de ≈ R$ 190.000,00, atingindo valores de até R$ 700.000,00. Este tipo de desequilíbrio e a discrepância considerável dos *outliers* em relação ao conjunto de dados, pode enviesar os métodos de MD, logo que sua ordem de grandeza influencia diretamente em medidas de distância. Por esse motivo, é necessária a realização da tarefa de pré-processamento descrita na seção 4.3.2 deste trabalho.



(a) (b)

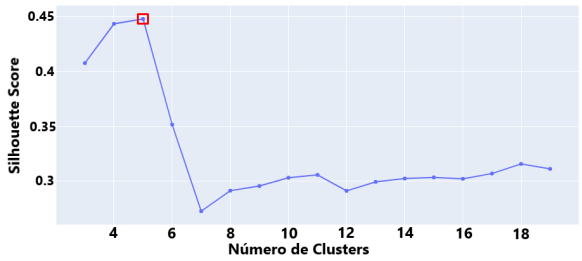
Figura 13: Gráficos *boxplot* e histograma do atributo “Valor total dos bens familiares”.

Por meio dos gráficos apresentados nessa seção foi possível entender algumas características de diversos atributos da base de dados. Entretanto, como pode ser observado na matriz de correlação (Figura 8), não é possível estabelecer conexão clara entre os atributos, evidenciando assim a necessidade da aplicação de métodos de *Data Mining* para obtenção de informações não triviais ao conjunto de dados. Seus resultados são descritos e discutidos nas seções seguintes.

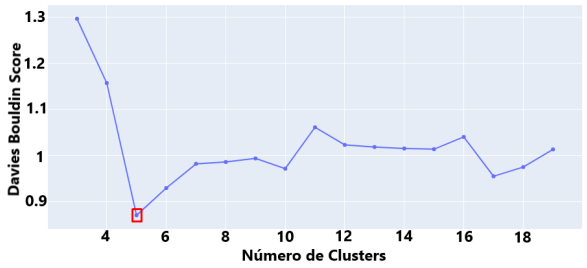
5.2 Identificação de Grupos

Para identificação dos grupos foram utilizados cinco métodos de aprendizado não-supervisionado, o K-Means, Agglomerative Clustering e Birch para o conjunto de dados numéricos (**6Num**) e os métodos K-Prototypes e K-Medoids para o conjunto de dados mistos (**6Num3Cat**). Para cada método há uma necessidade de determinar certos parâmetros que influenciam no resultado final do processo.

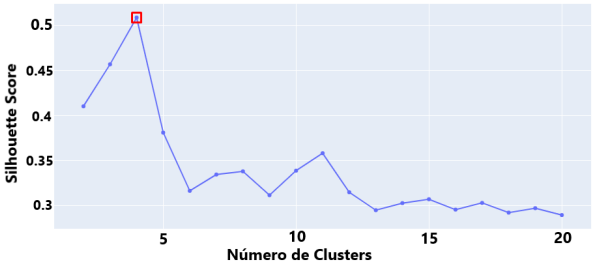
Para a escolha dos melhores parâmetros em relação a cada um dos métodos, foram executados testes variando principalmente o número de *clusters*, e em seguida foram construídos os gráficos referentes a avaliação dos índices. Estes testes foram realizados para os algoritmos K-Means, Agglomerative e Birch para os índices Silhouette, Davies-Bouldin e Dunn, para ilustrar esta etapa e embasar a escolha do número de *clusters* para cada método, a seguir são apresentados alguns dos gráficos:



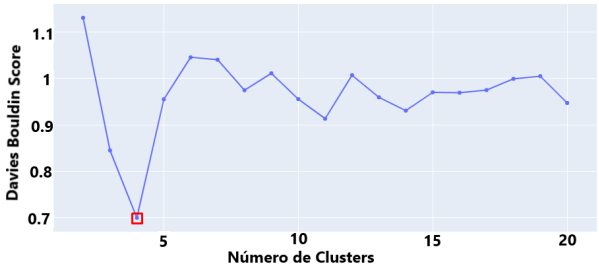
(a) Silhouette - Agglomerative (z-score)



(b) Davies-Bouldin - Agglomerative (z-score)



(c) Silhouette - K-Means (MIN-MAX com pesos)



(d) Davies-Bouldin - K-Means (MIN-MAX com pesos)

Figura 14: Gráfico da relação entre os índices Silhouette e Davies-Bouldin e o número de *clusters* para os algoritmos Agglomerative com z-score e K-Means com MIN-MAX com pesos.

Analisando a Figura 14 nota-se que o algoritmo Agglomerative com os dados normalizados com z-score apresenta o melhor valor dos índices Silhouette e Davies-Bouldin para uma quantidade de *clusters* igual a 5. Já para o método K-Means com a normalização MIN-MAX com pesos, a quantidade de grupos ideal é o valor 4. O valor destes índices fundamentam a escolha dos parâmetros dos algoritmos K-Means, Agglomerative e Birch.

Os demais resultados para estes métodos são apresentados na Tabela 6, que contém os melhores valores encontrados para os índices de avaliação Silhouette, Davies-Bouldin e Dunn em relação ao parâmetro mais relevante dos algoritmos de agrupamento, que é o número de grupos (*clusters*). Para cada método também foi avaliado as três normalizações consideradas neste trabalho: MIN-MAX (**MM**), MIN-MAX com peso (**MMp**) e z-score (**ZS**). O procedimento para obter cada pré-processamento está descrito na seção 4.3.2. Os melhores valores de cada índice estão destacados em negrito.

Tabela 6: Resultado dos índices de similaridade em relação a quantidade de *clusters*.

Algoritmo	**K-Means**			**Agglomerative**			**Birch**		
Pré-processamento	**MM**	**MMp**	**ZS**	**MM**	**MMp**	**ZS**	**MM**	**MMp**	**ZS**
Silhouette	0.493	0.509	0.359	0.532	0.509	0.448	0.523	0.509	**0.542**
Davies-Bouldin	0.757	0.699	**0.362**	0.702	0.693	0.870	0.708	0.691	0.827
Dunn	**0.869**	0.523	0.052	0.336	0.524	0.090	0.709	0.573	0.271
Número de *Clusters*	9	4	7	8	4	5	4	4	3

O número de *clusters*, mesmo sendo o parâmetro mais importante, não é o único com possibilidade de ajuste nos algoritmos. Durante o experimento tentou-se de forma empírica melhorar os resultado dos índices de similaridade a partir da alteração dos demais parâmetros de cada método. Contudo, os melhores resultados foram de fato obtidos com os valores *default*. Com exceção do algoritmo Birch, que apresentou os melhores resultados com o parâmetro *threshold* igual a 0.1, o qual foi considerado em todos os testes.

Mesmo com o auxílio dos índices, é complexo avaliar a qualidade de um agrupamento. Tendo como exemplo o algoritmo Birch com o pré-processamento z-score, que obteve o melhor resultado para o índice Silhouette. Quando observa-se o gráfico de distribuição dos indivíduos em relação aos grupos, presente na Figura 15, nota-se uma classificação desequilibrada, e que pode não ser útil quando se pensa no ponto de vista do gestor acadêmico precisar distribuir limitadas bolsas para um grupo de discentes.

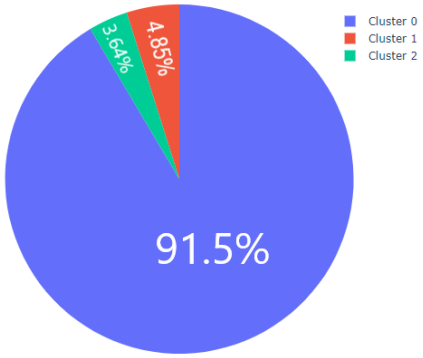
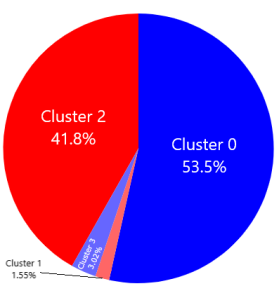
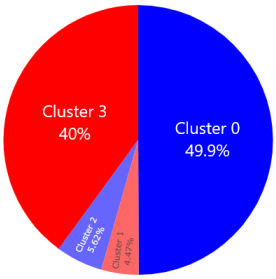


Figura 15: Gráfico da distribuição dos indivíduos pertencentes a cada grupo (*cluster*) para o Birch com normalização ZS.

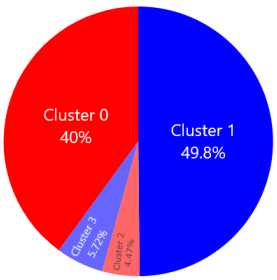
Observando outras abordagem é possível notar que a distribuição possui relacionamento direto com a quantidade de *clusters*, independente do algoritmo de agrupamento utilizado. A Figura 16 ilustra na parte superior os gráficos de distribuição para as abordagens Birch (MM), Birch (MMp) e Agglomerative (MMp), que possuem 4 grupos como resultado. Já as abordagens K-Means (MM), K-Means (ZS) e Agglomerative (MM), que apresentaram respectivamente 9, 7 e 8 grupos, são apresentadas na parte inferior da figura.



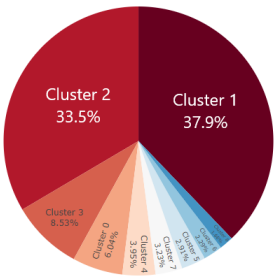
(a) Birch (MM)



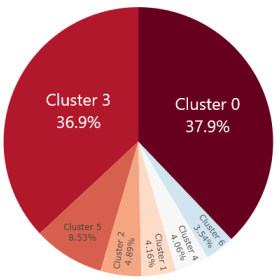
(b) Birch (MMp)



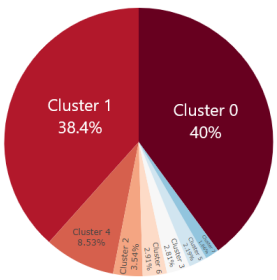
(c) Agglomerative (MMp)



(d) K-Means (MM)



(e) K-Means (ZS)



(f) Agglomerative (MM)

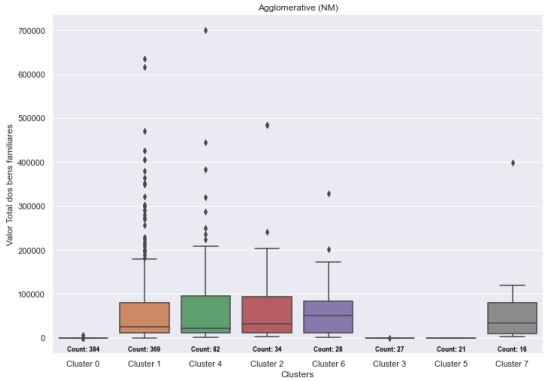
Figura 16: Distribuição dos indivíduos nos grupos para diferentes métodos para dados numéricos.

É evidente que a quantidade similar de grupos apresentou uma distribuição igualmente similar para qualquer que tenha sido a abordagem e normalização utilizada. É importante notar que no geral a distribuição se concentrou em dois grupos, tanto para as abordagens com quantidade de *clusters* menores quanto para quantidades maiores, o que pode indicar uma prevalência de dois grupos bem caracterizados no conjunto de dados. Entretanto, apenas com essa análise

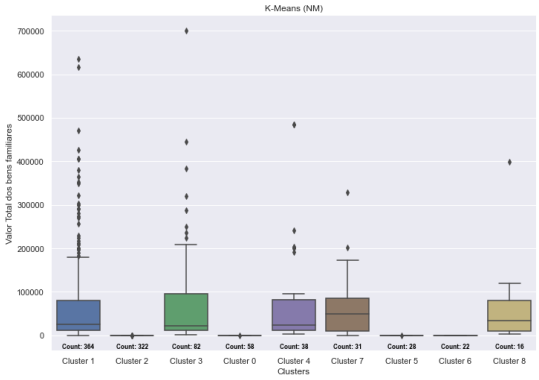
não é possível afirmar que os indivíduos de cada grupos possuem características semelhantes. Uma das maneiras de obter as características dos indivíduos de uma classe é a utilização de um classificador que forneça as regras de classificação, esta etapa é descrita a posteriori nesta seção.

Considerando os gráficos *boxplot* da Figura 17, que apresentam o valor da distribuição dos indivíduos de cada grupo para o atributo "Valor Total dos Bens Familiares", para os métodos Agglomerative (MM), K-Means (MM), Agglomerative (MMp) e K-Means (MMp), é possível notar que em todas as abordagens um dos grupos com grande quantidade de indivíduos possui sua média muito próxima ou igual ao valor R$ 0,00.

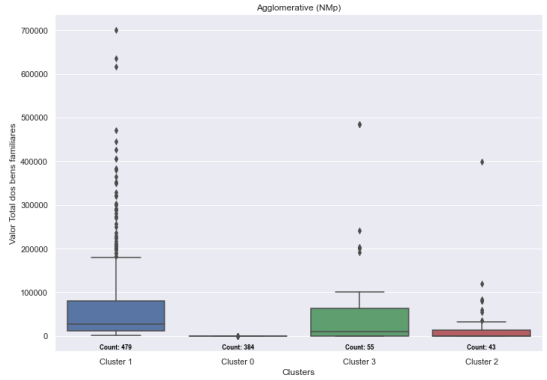
O *Cluster* 0 do método Agglomerative (MM) e Agglomerative (MMp) ambos com 384 indivíduos, o *Cluster* 2 do método K-Means (MM) com 322 registros e o *Cluster* 3 do método K-Means (MMp) com 382 indivíduos, apresentaram discentes com uma média de bens patrimoniais no valor de R$ 0,00 desconsiderando os *outliers*. Isto pode significar uma disposição dos métodos de encontrar um grupo com uma quantidade significativa de discentes que não apresentam patrimônio familiar.



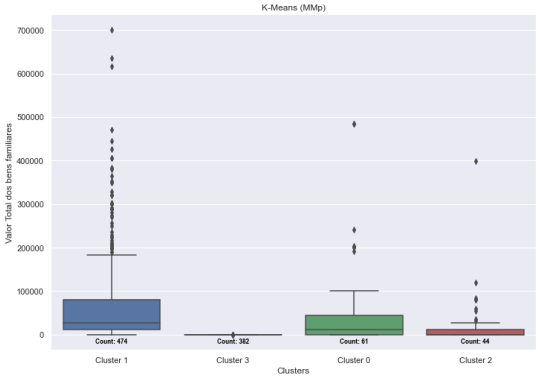
(a) Agglomerative (MM)



(b) K-Means (MM)



(c) Agglomerative (MMp)

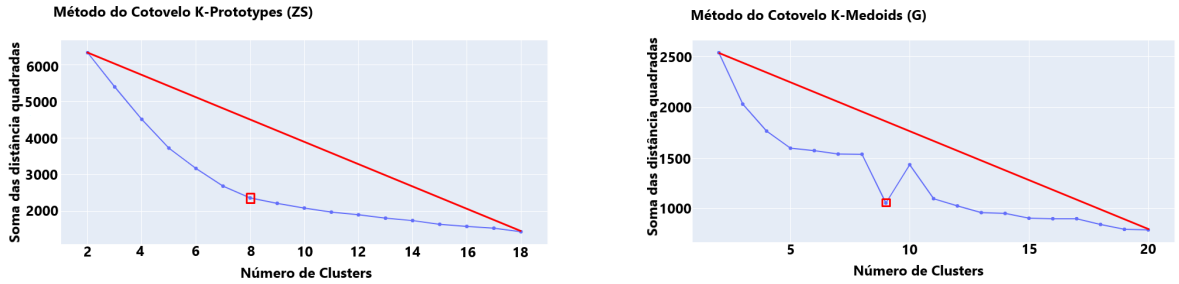


(d) K-Means (MMp)

Figura 17: Distribuição dos indivíduos em cada grupo para o atributo "Valor Total dos Bens Familiares".

Para abordar uma outra perspectiva quanto ao agrupamento, foi realizada a execução dos métodos K-Prototypes e K-Medoids no conjunto de dados **6Num3Cat**. Para o método K-Medoids foi utilizado a medida de distância Gower (**G**), que gera uma matriz de distância entre todos os atributos do conjunto de dados, tanto numéricos quanto categóricos. Foi testado também esta abordagem com a aplicação de peso dois após a geração da matriz para os atributos "Renda per capita" e "Despesas per Capita", este nomeado como Gower com peso (**Gp**). Já o método K-Prototypes, por fazer um agrupamento dos dados categóricos e numéricos separadamente para depois fazer a fusão, foi utilizado os mesmos pré-processamentos utilizados nos métodos numéricos, o MIN-MAX (**MM**), o MIN-MAX com peso (**MMp**) e o z-score (**ZS**). Este pré-processamento é realizado apenas nos atributos numéricos.

O índice utilizado para identificar os melhores resultados foi o índice da soma das distâncias quadradas, conhecido como Método do Cotovelo (*Elbow Method*). Da mesma maneira executada nos métodos numéricos, os métodos K-Prototypes e K-Medoids foram testados com diferentes parâmetros buscando encontrar os melhores resultados para o índice. Entretanto, seus melhores resultados foram observados ao se considerar os valores *default* dos métodos, necessitando a alteração apenas da quantidade de grupos. Para ilustrar a escolha do número de *clusters*, alguns dos resultados são apresentados na Figura 18.



(a) K-Prototypes (ZS) (b) K-Medoids (G)

Figura 18: Exemplos da aplicação do Método do Cotovelo para identificação da quantidade ideal de *clusters* para os algoritmos K-Prototypes (ZS) e K-Medoids (G).

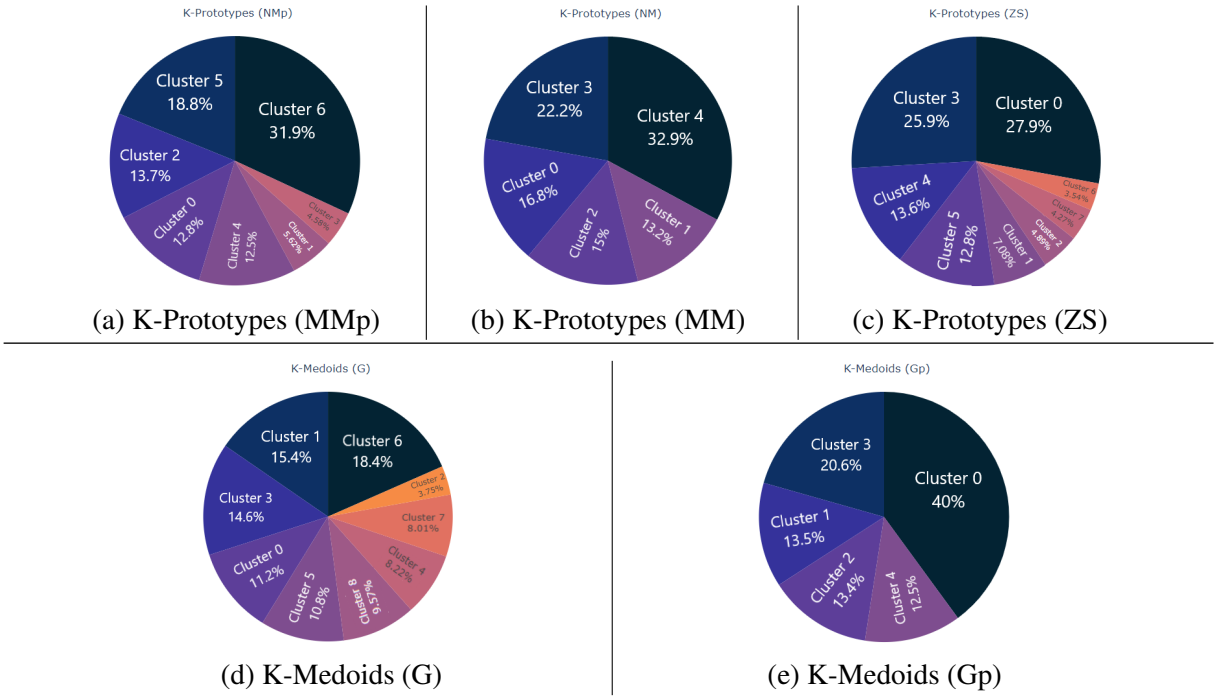
O melhor valor para o Método do Cotovelo é a quantidade de *clusters* que se encontra mais distantes da linha traçada entre o ponto 1 e o ponto n. Os demais resultados encontrados para os métodos K-Prototypes e K-medoids, com seus respectivos pré-processamentos, são listados na Tabela 7.

Tabela 7: Melhor valor para o Método do Cotovelo em relação a quantidade de *clusters*.

Algoritmo	**K-Prototypes**			**K-Medoids**	
Pré-processamento	**MM**	**MMp**	**ZS**	**G**	**Gp**
Número de *Clusters*	5	7	8	9	5

Diferente do resultado dos métodos puramente numéricos, a distribuição dos indivíduos em seus respectivos *clusters* não apresentou simetria notável entre as abordagens, como é apre-

sentado na Figura 19. Observa-se que a distribuição ocorre de maneira mais equilibrado nos métodos que utilizam dados mistos, onde cada abordagem apresenta um resultado mais singular. Verifica-se também que na maioria dos resultados, tanto para métodos exclusivamente numéricos quanto para dados mistos, dois dos grupos somados representam mais de 50% do conjunto de dados e por vezes até 90%. Isto pode ser um indicativo de que os indivíduos apresentam características muito semelhantes entre si, tornando o conjunto de dados homogêneo e complexo de se desmembrar. Este é um entendimento lógico, pois o conjunto de dados têm origem em um requerimento voluntário de auxílio financeiro. Portanto espera-se que os indivíduos apresentem características socioeconômicas similares.

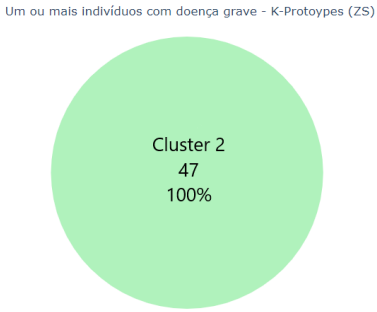


(a) K-Prototypes (MMp) (b) K-Prototypes (MM) (c) K-Prototypes (ZS)

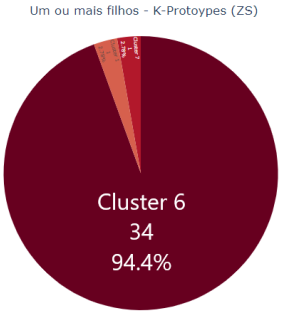
(d) K-Medoids (G) (e) K-Medoids (Gp)

Figura 19: Distribuição dos indivíduos nos grupos para os métodos K-Prototypes e K-Medoids com diferentes pré-processamentos.

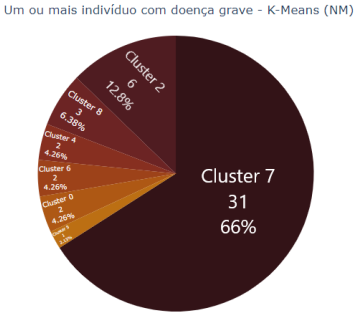
Outro ponto observado, foi a tendência de alguns métodos de separar grupos exclusivamente por um atributo. Isto é ilustrado na Figura 20, na qual a Figura 20a referente ao método K-Prototypes (ZS) representa a distribuição entre os grupos quando o atributo "Quantidade de individuos com doença grave no grupo familiar" é maior ou igual a 1, ou seja, se algum membro da família apresenta alguma doença grave. O gráfico demonstra que todos os 47 indivíduos do conjunto de dados que possuem parentes, ou ele próprio, com doença grave estão alocados no *Cluster 2*. Complementando o gráfico, a quantidade de indivíduos no *Cluster 2* é de exatamente 47, portanto é possível afirmar que o grupo dois foi separado exclusivamente pelo atributo antes citado.



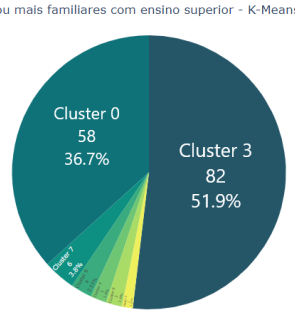
(a) Um ou mais indivíduos com doença grave - K-Prototypes (ZS)



(b) Um ou mais filhos - K-Prototypes (ZS)



(c) Um ou mais indivíduos com doença grave - K-Means (MM)



(d) Um ou mais familiares com superior completo ou pós - K-Means (MM)

Figura 20: Gráficos de distribuição em grupos dos atributos “Quantidade de individuos com doença grave no grupo familiar” e “Quantos filhos o solicitante possui?” para valores acima de zero, para o método K-Prototypes (ZS) e K-Means(MM).

Isto acontece também no gráfico da Figura 20b, onde o *Cluster 6* apresenta 34 indivíduos com um ou mais filhos, exatamente a quantidade de registros total do *Cluster 6*. E repete-se para a Figura 20c, onde o total de registros do *Cluster 7* é de 31 discentes. Isto indica que alguns métodos separaram certos grupos exclusivamente por conta de certa característica. Já para a Figura 20d no qual os *Cluster 0* e *Cluster 3* possuem respectivamente os exatos 58 e 82 registos, foi constatado que em relação aos outros atributos o que diferencia os dois grupos é o atributo “Valor Total dos Bens Familiares”, apresentando uma média de R$ 0,00 para o *cluster 0* e R$ 75.000,00 para o *cluster 3*.

Por conta do conjunto de dados apresentar uma quantidade de aproximadamente mil registros, o tempo de execução dos métodos não-supervisionados foi baixo. O método mais veloz foi o Birch com média de 0.082 segundos, e o mais lento o K-Prototypes com média de 8.325 segundos. Tendo em vista o objetivo da pesquisa estes valores são irrisórios, logo que a celeridade dos métodos é um aspecto pouco relevante.

Para descobrir outros conhecimentos sobre os grupos elaborados pelos métodos não-supervisionados, é preciso executar a próxima etapa no fluxo da modelagem, que consiste na execução do classificador que identificará as regras implícitas a cada grupo. O classificador utilizado para criação da Árvore de Decisão foi o J48 implementado no *software* WEKA.

A análise pode ser feita a partir da observação da Árvore de Decisão e da tabela com as regras de decisão do classificador para cada grupo. As Árvores de Decisão subsequentes apresentam as regras para alcançar cada nó folha, que representa um *cluster* do conjunto de dados, no qual é apresentado o número de indivíduos classificados corretamente, e separado por uma barra o número de indivíduos classificados incorretamente naquele *cluster*. Já as tabela apresentam as mesmas regras de decisão, porém com a porcentagem de indivíduos em cada grupo com relação a todo o conjunto de dados. São apresentados os melhores métodos para dados numéricos segundo os índices de similaridade, são eles: K-Means (MM), K-Means (ZS) e Birch (ZS). E para os métodos para dados mistos foram escolhidos os que apresentaram o maior tamanho da árvore de decisão, para analisar uma possível complexidade maior nas regras, que são: K-Prototypes (ZS) e K-Medoids (G).

Para uma melhor apresentação nas tabelas o nome dos atributos foi abreviado da seguinte forma: **[REN]**: Renda per capita; **[DES]**: Despesas per capita; **[FSC]**: Familiares com Superior Completo ou Pós; **[VBF]**: Valor Total dos Bens Familiares; **[DGF]**: Quantidade de indivíduos com doença grave no grupo familiar; **[QFL]**: Quantos filhos o solicitante possui; **[PES]**: Qual sua Procedência Escolar?; **[MAL]**: Qual a situação da Moradia do Aluno?; **[MFA]**: Qual a situação da Moradia do Grupo Familiar?

O valor dos atributos categóricos também foi abreviado, da seguinte maneira: **[AC]**: Aluguel com Colegas; **[FT]**: Família/Terceiros; **[AFS]**: Aluguel/ Financiamento Sozinho; **[PQ]**: Próprio Quitado; **[HR]**:Herança; **[AL]**: Alugada; **[PP]**: Própria em Pagamento; **[P]**: Particular; **[PB]**: Particular Bolsa >50%; **[Pb]**: Particular Bolsa <50%; **[F]**: Filantrópica; **[PU]**: Pública.

A Figura 21 e a Tabela 8 apresentam as regras e a porcentagem de indivíduos em cada grupo para o conjunto de dados 6Num com o pré-processamento z-score e o método de agrupamento K-Means que foi o melhor resultado para o índice Davies-Bouldin.

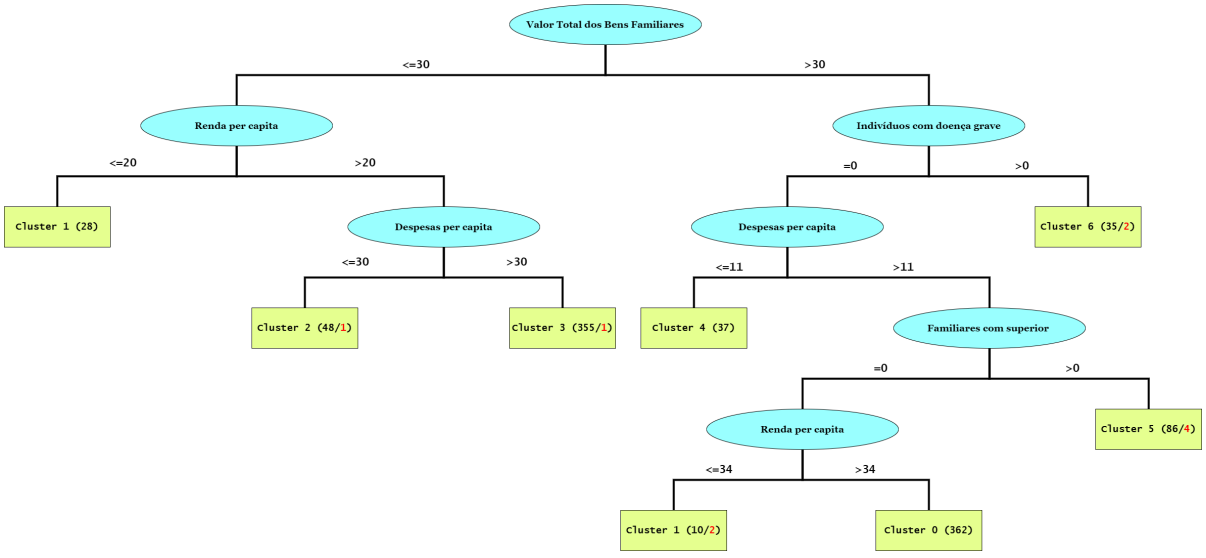


Figura 21: Árvore de Decisão obtidas pelo método K-Means (ZS) para o conjunto de dados 6Num.

Tabela 8: Regras de classificação obtidas pelo método K-Means (ZS) para o conjunto de dados 6Num.

Grupos	**Distribuição**	**Regras de Decisão**
Grupo 0	38%	Se VBF>30 & DGF=0 & DES>11 & FSC=0 & REN>34
Grupo 1	4%	Se VBF<=30 & REN<=20 **ou** Se VBF>30 & DGF=0 & DES>11 & FSC=0 & REN<=34
Grupo 2	5%	Se VBF<=30 & REN>20 & DES<=30
Grupo 3	37%	Se VBF<=30 & REN>20 & DES>30
Grupo 4	4%	Se VBF>30 & DGF=0 & DES<=11
Grupo 5	9%	Se VBF>30 & DGF=0 & DES>11 & FSC>0
Grupo 6	3%	Se VBF>30 & DGF>0

A Figura 22 e a Tabela 9 apresentam os resultados para o conjunto de dados 6Num com o pré-processamento MIN-MAX com peso e o método de agrupamento K-Means que obteve o melhor resultado para o índice Dunn.

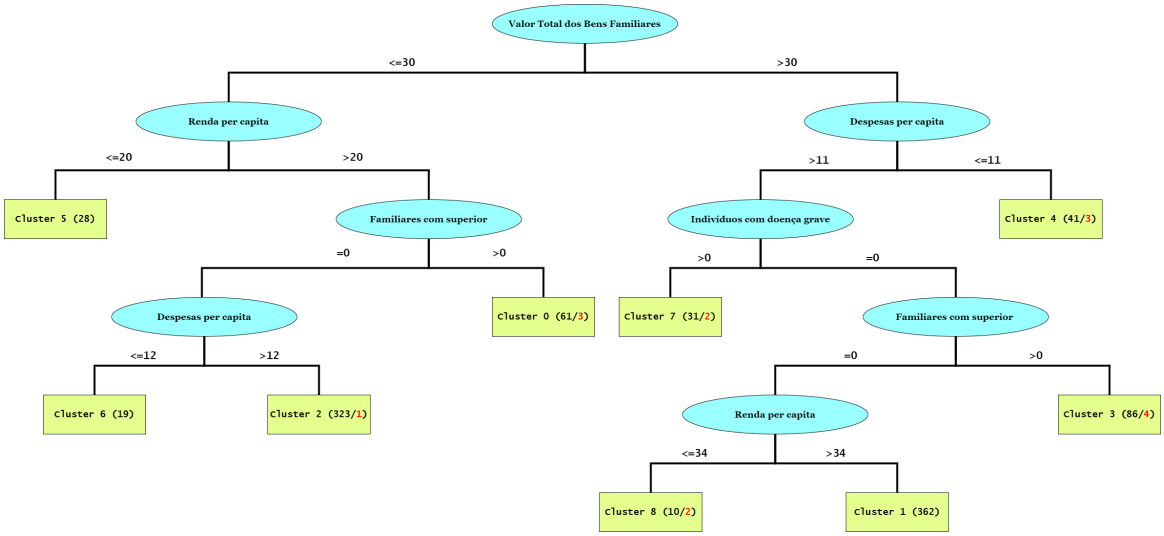


Figura 22: Árvore de Decisão obtidas pelo método K-Means (MM) para o conjunto de dados 6Num.

Tabela 9: Regras de classificação obtidas pelo método K-Means (MM) para o conjunto de dados 6Num.

Grupos	**Distribuição**	**Regras de Decisão**
Grupo 0	6%	Se VBF<=30 & REN>20 & FSC>0
Grupo 1	38%	Se VBF>30 & DES>11 & DGF=0 & FSC=0 & REN>34
Grupo 2	33%	Se VBF<=30 & REN>20 & FSC=0 & DES>12
Grupo 3	9%	Se VBF>30 & DES>11 & DGF=0 & FSC>0
Grupo 4	4%	Se VBF>30 & DES<=11
Grupo 5	3%	Se VBF<=30 & REN<=20
Grupo 6	2%	Se VBF<=30 & REN>20 & FSC=0 & DES<=12
Grupo 7	3%	Se VBF>30 & DES>11 & DGF>0
Grupo 8	2%	Se VBF>30 & DES>11 & DGF=0 & FSC=0 & REN<=34

A Figura 23 e a Tabela 10 os resultados para o conjunto de dados 6Num com o pré-processamento z-score e o método Birch que foi o melhor resultado para o índice Silhouette.

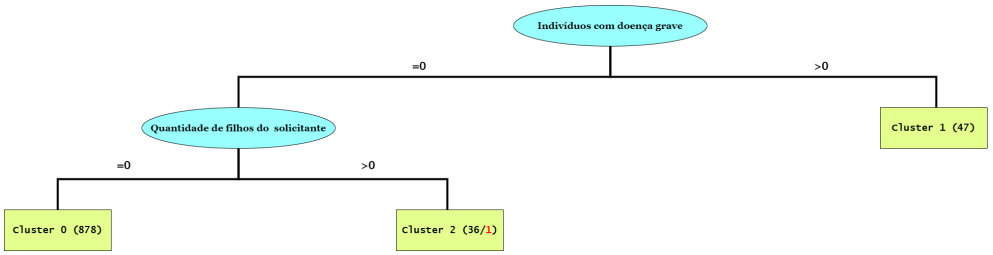


Figura 23: Árvore de Decisão obtidas pelo método Birch (ZS) para o conjunto de dados 6Num.

Tabela 10: Regras de classificação obtidas pelo método Birch (ZS) para o conjunto de dados 6Num.

Grupos	**Distribuição**	**Regras de Decisão**
Grupo 0	91%	Se DGF=0 & QFL=0
Grupo 1	5%	Se DGF>0
Grupo 2	4%	Se DGF=0 & QFL>0

Como é possível observar, a abordagem Birch (ZS) utilizou predominantemente apenas dois atributos para realizar a divisão dos grupos, foram eles: "Quantidade de indivíduos com-doença grave no grupo familiar" e "Quantos filhos o solicitante possui". Somado a distribuição desequilibrada, conclui-se que esta abordagem não apresentou resultado satisfatório para os objetivos da pesquisa, mesmo tendo o melhor valor de acordo com o índice Silhouette.

Já as regras para as abordagens K-Means (ZS) e K-Means (MM), apresentaram resultados similares, tendo em vista que: possuem dois grupos com valor em torno de 35% da amostra; que os atributos mais relevantes foram "Valor Total dos Bens Familiares", "Renda per capita" e "Despesas per capita", logo que são os que mais se repetem no conjunto de regras; e que suas condições oscilam na casa de algumas dezenas. O que leva a crer que, apesar da quantidade de grupos diferentes e do pré-processamento distinto, as abordagens realizaram a divisão dos grupos de forma semelhante.

As condições presentes nas regras para os atributos *VBF*, *REN* e *DES* variaram entre os valores R$ 0,00 e R$ 34,00 estes valores são baixos considerando as médias destes atributos: VBF_{mean} = R$ 36.709,00 ; REN_{mean} = R$ 548,00 ; DES_{mean} = R$ 154,00. Uma provável explicação é que muitos discentes apresentem valores próximos a R$ 0,00, principalmente para o campo referente ao valor dos bens familiares. Como exemplo, os dois grupos com mais registros destas duas abordagens, o *Grupo 0* e *Grupo 3* do método K-Means (ZS), e o *Grupo 1* e *Grupo 2* do método K-Means (MM), no qual suas principais diferenças entre os grupos, para ambos os métodos, é o valor patrimonial variando de menor ou igual a R$ 30,00 e maior que R$ 30,00. Isto pode ser lido na prática como sendo R$ 0,00 ou maior que R$ 0,00 logo que o valor R$ 30,00 é irrisório em relação média do atributo, que é de R$ 36.709,00. Portanto conclui-se

que para os métodos K-Means (ZS) e K-Means (MM), o atributo mais significativo foi o atributo *VBF*, logo que mais de 70% do conjunto de dados foi dividido tendo este atributo como variável principal. Os demais 30% foram divididos considerando principalmente subconjuntos de dados com características peculiares, como Renda per capita e Despesas per Capita próximas a R$ 0,00 sendo um comportamento incomum considerando as necessidades básicas de um ser humano.

Para os métodos que utilizaram dados mistos suas Árvores de Decisão se mostraram grandes para apresentação. Portanto suas regras foram dispostas apenas em formato de tabela. Os resultados para o conjunto de dados 6Num3Cat com o pré-processamento z-score e com o método de agrupamento K-Prototypes estão presentes na Tabela 11. E para o conjunto de dados 6Num3Cat com a medida de distância Gower e com o método K-Medoids são apresentados na Tabela 12.

Tabela 11: Regras de classificação obtidas pelo método K-Prototypes (ZS) para o conjunto de dados 6Num3Cat.

Grupos	**Distribuição**	**Regras de Decisão**
Grupo 0	28%	Se DGF=0 & FSC=0 & REN<=729 & QFL=0 & DES<=91 & VBF<=196000 & MAL=AC & REN>340 **ou** Se DGF=0 & FSC=0 & REN<=729 & QFL=0 & DES>91 & DES<=333 & VBF<=80000 & MAL=AC **ou** Se DGF=0 & FSC=0 & REN<=729 & QFL=0 & DES>136 & DES<=333 & VBF<=80000 & MAL=FT **ou** Se DGF=0 & FSC=0 & REN<=729 & QFL=0 & DES>141 & DES<=333 & VBF<=80000 & MAL=AFS **ou** Se DGF=0 & FSC=0 & REN<=618 & QFL=0 & DES>91 & DES<=333 & VBF>80000
Grupo 1	7%	Se DGF=0 & FSC=0 & REN>729 & VBF<=190290 & DES>371 **ou** Se DGF=0 & FSC>0 & DES>462
Grupo 2	5%	Se DGF>0
Grupo 3	26%	Se DGF=0 & FSC=0 & REN<=340 & QFL=0 & DES<=91 & VBF<=196000 & MAL=AC **ou** Se DGF=0 & FSC=0 & REN<=729 & QFL=0 & DES<=91 & VBF<=196000 & MAL=FT **ou** Se DGF=0 & FSC=0 & REN<=729 & QFL=0 & DES<=91 & VBF<=196000 & MAL=AFS **ou** Se DGF=0 & FSC=0 & REN<=729 & QFL=0 & DES<=91 & VBF<=196000 & MAL=PQ **ou** Se DGF=0 & FSC=0 & REN<=729 & QFL=0 & DES>91 & DES<=136 & VBF<=80000 & MAL=FT **ou** Se DGF=0 & FSC=0 & REN<=729 & QFL=0 & DES>91 & DES<=333 & VBF<=80000 & MAL=PQ **ou** Se DGF=0 & FSC=0 & REN<=729 & QFL=0 & DES>91 & DES<=141 & VBF<=80000 & MAL=AFS **ou** Se DGF=0 & FSC=0 & REN>729 & VBF<=190290 & DES<=76 & REN<=770
Grupo 4	14%	Se DGF=0 & FSC=0 & REN>729 & VBF<=190290 & DES<=371 & REN<=770 & DES>76 **ou** Se DGF=0 & FSC=0 & REN>729 & VBF<=190290 & DES<=371 & REN>770
Grupo 5	13%	Se DGF=0 & FSC>0 & DES<=462 & VBF <=204619
Grupo 6	3%	Se DGF=0 & FSC=0 & REN<=729 & QFL=1
Grupo 7	4%	Se DGF=0 & FSC=0 & REN<=729 & QFL=0 & DES<=91 & VBF>196000 **ou** Se DGF=0 & FSC=0 & REN<=729 & QFL=0 & DES>91 & DES<=333 & VBF>80000 & REN>618 **ou** Se DGF=0 & FSC=0 & REN>729 & VBF>190290 **ou** Se DGF=0 & FSC>0 & DES<=462 & VBF>204619

Tabela 12: Regras de classificação obtidas pelo método K-Medoids (com a medida de distância Gower) para o conjunto de dados 6Num3Cat.

Grupos	**Distribuição**	**Regras de Decisão**
Grupo 0	11%	Se MFA=HR & MAL=AC
Grupo 1	15%	Se MFA=PQ & MAL=AC & FSC=0 & PES=PU & VBF<=208763
Grupo 2	4%	Se MFA=PP & PES=PU & MAL=AC **ou** Se MFA=PP & PES=PU & MAL=AFS **ou** Se MFA=PP & PES=PU & MAL=PQ **ou** Se MFA=PP & PES=F **ou** Se MFA=PP & PES=P
Grupo 3	15%	Se MFA=HR & MAL=FT **ou** Se MFA=HR & MAL=AFS **ou** Se MFA=HR & MAL=PQ
Grupo 4	8%	Se MFA=PP & PES=PB **ou** Se MFA=PP & PES=Pb
Grupo 5	11%	Se MFA=AL & MAL=FT **ou** Se MFA=AL & MAL=AFS **ou** Se MFA=AL & MAL=PQ **ou** Se MFA=PP & PES=PU & MAL=FT
Grupo 6	18%	Se MFA=PQ & MAL=FT **ou** Se MFA=PQ & MAL=AFS & REN<=659 & VBF<=11100 **ou** Se MFA=PQ & MAL=PQ
Grupo 7	8%	Se MFA=AL & MAL=AC
Grupo 8	10%	Se MFA=PQ & MAL=AC & FSC=0 & PES=PU & VBF>208763 **ou** Se MFA=PQ & MAL=AC & FSC=0 & PES=PB **ou** Se MFA=PQ & MAL=AC & FSC=0 & PES=F **ou** Se MFA=PQ & MAL=AC & FSC=0 & PES=Pb **ou** Se MFA=PQ & MAL=AC & FSC=0 & PES=P **ou** Se MFA=PQ & MAL=AC & FSC>0 **ou** Se MFA=PQ & MAL=AFS & REN<=659 & VBF>11100 **ou** Se MFA=PQ & MAL=AFS & REN>659

Na Tabela 12 observa-se uma predominância dos atributos *MFA*, *PES* e *MAL*, que correspondem aos três atributos categóricos presentes no conjunto de dados 6Num3Cat, que se referem respectivamente aos campos: “Qual a situação da Moradia do Grupo Familiar?”; “Qual sua Procedência Escolar?” e “Qual a situação da Moradia do Aluno?”. Isto indica que o método atribuiu uma importância excessiva sobre estes atributos. Pode-se imaginar que este resultado se deu por conta da matriz de distância construída pelo método Gower, possivelmente sendo necessário um pré-processamento no conjunto de dados numéricos antes da geração da matriz, imaginando assim que os dados numéricos se tornem mais relevantes para o algoritmo de agrupamento.

Já a Tabela 11 traz consigo resultados complexos de se avaliar, logo que não possui desequilíbrio exagerado entre os grupos, suas regras apresentam a maior parte dos atributos do conjunto de dados e as condições possuem valores razoáveis quanto as médias dos atributos, sendo este o resultado mais promissor do ponto de vista da heterogeneidade dos grupos. A complexidade envolvida nas regras apresentadas não necessariamente significam um melhor resultado para o objetivo final da pesquisa, tendo em vista que os dados possuem características similares entre si. Entretanto, a heterogeneidade encontrada pode ser um ponto interessante considerando a necessidade de uma distinção não trivial entre os grupos.

Ao fim, tem-se duas abordagens que não apresentaram resultados satisfatórios por separarem os grupos de maneira trivial e desconsiderar atributos importantes como despesa e renda, são elas Birch (SZ) e K-Medoids (G). Já os métodos K-Means (ZS) e K-Means (MM), apli-

cados exclusivamente a dados numéricos, apresentaram os melhores valores para os índices de similaridade Davies-Bouldin e Dunn, respectivamente; exibiram regras de decisão que, embora simples, podem ter efeito prático para o gestor acadêmico; distribuíram-se em dois grupos principais com aproximadamente 35% da amostra para cada; e seu atributo mais significativo foi o *VBF*. E o método K-Prototypes que apresentou o resultado mais heterogêneo entre as abordagens, com uma distribuição pouco desequilibrada e considerando a maior parte dos atributos em seu conjunto de regras de decisão, tendo assim regras mais complexas em relação aos outros métodos.

5.3 Classificação de Novos Solicitantes

Foi por conta da classificação supervisionada que se determinou as regras para a distribuição dos indivíduos em seus respectivos grupos. Entretanto essas regras são geradas a partir de um conjunto de treinamento, e é por meio das regras que são classificados os demais solicitantes. A Tabela 13 apresenta o resultado do classificador mediante aos índices de validação apresentados na seção 4.5, e tendo como entrada todos os resultados da etapa de agrupamento. A seguir os valores são apresentados para cada tipo de pré-processamento e método de agrupamento adotado.

Observa-se que todos os resultados referentes aos índices Kappa, Acurácia Balanceada e Média AUC apresentaram valores próximos ao valor ideal. Isto indica que o conjunto de regras encontrado pelo classificador para a identificação de cada grupo, representa de forma autêntica as características dos indivíduos daquele grupo, possibilitando uma classificação com baixo erro de um novo indivíduo. Porém, também pode indicar que as técnicas de pré-processamento juntamente aos métodos de agrupamento estão dividindo grupos de maneira trivial, a medida que as regras para divisão dos indivíduos é de baixa complexidade, principalmente para os casos onde o Tamanho da Árvore é pequeno. Isto pode ser notado pelos valores apresentados para o campo Tamanho da Árvore dos métodos que utilizaram dados exclusivamente numéricos em contraponto aos métodos para dados mistos, que resultaram em tamanhos de árvore maiores. Entretanto, este comportamento é esperado devido ao aumento da quantidade de atributos do conjunto 6Num para o conjunto 6Num3Cat.

Tabela 13: Resultado dos índices de validação para a Árvore de Decisão do algoritmo J48.

Agrupamento	**K-Means**			**Agglomerative**			**Birch**			**K-Prototypes**			**K-Medoids**	
Pré-Processamento	**MM**	**MMp**	**ZS**	**MM**	**MMp**	**ZS**	**MM**	**MMp**	**ZS**	**MM**	**MMp**	**ZS**	**G**	**Gp**
Kappa	0.964	0.995	0.981	0.965	0.995	0.963	**0.998**	0.996	0.993	0.942	0.904	0.867	0.851	0.872
Acurácia Balanceada	0.962	0.989	0.962	**0.999**	0.995	**0.999**	0.894	0.994	0.930	0.941	0.892	0.878	0.829	0.871
AUC	0.992	0.997	0.994	0.996	0.997	0.992	**0.999**	0.998	**0.999**	0.990	0.975	0.972	0.973	0.973
Tempo de execução (*s*)	0.02	0.01	0.02	0.02	0.01	0.01	0.01	0.01	0.01	0.03	0.04	0.05	0.05	0.03
Tamanho da Árvore	17	7	15	15	7	11	7	7	5	27	35	45	41	25

6 Conclusões e Trabalhos Futuros

Em um passado recente da história brasileira, políticas de acesso a universidade foram ampliadas, sobretudo para a população socioeconomicamente mais vulnerável. Esta situação trás consigo uma necessidade maior da efetividade de programas de assistência estudantil, logo que a situação financeira de um indivíduo ou grupo familiar têm impacto direto em índices acadêmicos de rendimento e de evasão, além de prejuízos a saúde mental e ao bem estar social do discente. Este trabalho buscou avaliar métodos de Mineração de Dados Educacionais mediantes a índices e percepções estatísticas, em dados socioeconômicos de requerentes de auxílio permanência da UNIFEI no ano de 2018, para validar um modelo de caracterização de perfis que possa auxiliar o gestor acadêmico no processo de escolha dos bolsistas.

Foram testados os métodos de aprendizado não-supervisionado K-Means, Agglomerative Clustering e Birch para uma seleção de atributos exclusivamente numéricos e os métodos K-Prototypes e K-Medoids (com a medida de distância Gower) para uma seleção de atributos numéricos e categóricos. Este métodos foram executados com a intenção de encontrar grupos com diferentes perfis no conjunto de dados. A seguir, foi executado o método de aprendizado supervisionado J48 para a geração da Árvore de Decisão. Esta etapa gerou as regras de características de cada grupo e avaliou a classificação de novos indivíduos nos grupos pré-definidos.

Apesar de métodos como Birch com pré-processamento z-score e K-Medoids com a medida de distância Gower, não apresentarem resultados satisfatórios, logo que subdividiram grupos predominantemente com base em dois ou três atributos e ignoraram características cruciais para a análise como despesas e renda do grupo familiar, outros métodos como K-Means e K-Prototypes exibiram resultados promissores para o objetivo de classificar diferentes perfis de discentes, sendo observado suas distribuições e regras de características para cada grupo. O método K-Means tanto para o pré-processamento MIX-MAX quanto para o pré-processamento z-score obtiveram resultados relevantes quanto aos índices de similaridade e apresentaram um conjunto de regras coerentes quanto a distinção de perfis. Já o método K-Protoypes com o pré-processamento z-score nos dados numérico, apresentou os resultados mais heterogêneos e complexos quanto ao conjunto de regras, sua distribuição se mostrou mais simétrica, divergindo assim da maior parte dos métodos.

Devido a obtenção de resultados distintos de classificação e por ser um agente externo aos processos da gestão acadêmia, não foi possível apontar um modelo ideal para a classificação de diferentes perfis de requentes. Entretanto, é possível afirmar que métodos de Mineração de Dados Educacionais podem ser úteis no processo de tomada de decisão, logo que seus resultados são otimistas para com o agrupamento dos diferentes perfis socioeconômicos de discentes.

De modo a melhorar a qualidade do agrupamento sugere-se a realização de estudos em outras bases de dados, variando o ano ou até mesmo a universidade. Tendo em vista os melhores métodos apontados nessa pesquisa, K-Means e K-Prototypes, propõe-se que para um aprimoramento dos modelos, as abordagens sejam executados alterando: a seleção dos atributos, o pré-processamento utilizado, e os índices de similaridade adotados para identificar o melhor resultado. É sugerido também que a pesquisa ocorra de forma interdisciplinar, com profissionais da área da gestão acadêmica intervindo em questões como atributos mais significativos e avaliando resultados no decorrer da pesquisa. E que a DAE altere o formulário para recepção destes dados, de forma a limitar as respostas a cada questão, para que em futuras análises a limpeza dos dados não seja questão custosa quanto ao tempo.

Como trabalhos futuros, além das alterações supracitadas, sugere-se a execução da fase 6 do modelo CRISP-DM (Aplicação), e a comparação entre os grupos de bolsistas classificados manualmente pela equipe técnica competente e os resultados dos modelos de Mineração de Dados apresentados.

Referências

ALENCAR, M.; SANTOS, E.; NETTO, J. M. Identifying students with evasion risk using data mining. In: . [S.l.: s.n.], 2015.

BAKER, R. S. Data mining for education. *International Encyclopedia of Education*, in press.

BAKER, R. S. J. de; CARVALHO, A. M. J. B. de; ISOTANI, S. Mineração de dados educacionais: Oportunidades para o brasil. *Revista Brasileira de Informática na Educação*, v. 19, 2011.

BARADWAJ, B. K.; PAL, S. Mining educational data to analyze students' performance. *International Journal of Advanced Computer Science and Applications*, 2011.

BERENS, J. et al. Early detection of students at risk: Predicting student dropouts using administrative student data from german universities and machine learning methods. *Journal of Educational Data Mining*, v. 11, n. 3, p. 1–41, 2019.

BERRY, M. J. A.; LINOFF, G. S. *Data Mining Techniques*: For marketing, sales and customer relationship management. 2. ed. [S.l.]: Wiley, 2004. ISBN 0-471-47064-3.

BERRY, M. W.; MOHAMED, A.; YAP, B. W. *Supervised and Unsupervised Learning for Data Science*. [S.l.]: Springer, 2020. ISBN 978-3-030-22475-2.

BRASIL. Decreto nž 6.096, de 24 de abril de 2007 - programa de apoio a planos de reestruturação e expansão das universidades federais - reuni. *Diário Oficial da República Federativa do Brasil*, 2007.

BRASIL. Decreto nž 7.234, de 19 de julho de 2010 - programa nacional de assistência estudantil - pnaes. *Diário Oficial da República Federativa do Brasil*, 2010.

BRITO, W. M. de; SEMAAN, G. S.; BRITO, J. A. de M. Um algoritmo genÉtico para o problema dos k-medoids. *Brazilian Congress on Computational Intelligence*, 2011.

CERCHIARI, E. A. N. Saúde mental e qualidade de vida em estudantes universitários. 2004.

CHAPMAN, P. et al. *CRISP-DM 1.0*. 2000. Disponível em: <http://www.the-modeling-agency.com/crisp-dm.pdf>.

COHEN, J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, v. 20, n. 1, p. 37–46, 1960. Disponível em: <https://doi.org/10.1177-/001316446002000104>.

CORTES, S. da C.; PORCARO, R. M.; LIFSCHITZ, S. Mineração de dados - funcionalidades, técnicas e abordagens. 2002.

COSTA, E. et al. Mineração de dados educacionais: Conceitos, técnicas, ferramentas e aplicações. *Jornada de Atualização em Informática na Educação*, p. 1–29, 2012. ISSN 23167734.

DAMANIK, I. S. et al. Decision tree optimization in c4.5 algorithm using genetic algorithm. *Journal of Physics: Conference Series*, IOP Publishing, v. 1255, p. 012012, aug 2019. Disponível em: <https://doi.org/10.1088/1742-6596/1255/1/012012>.

DASH, M.; LIU, H.; YAO, J. Dimensionality reduction of unsupervised data. In: *Proceedings Ninth IEEE International Conference on Tools with Artificial Intelligence*. [S.l.: s.n.], 1997. p. 532–539.

DESGRAUPES, B. Clustering indices. *University Paris Ouest - Lab ModalX*, 2017.

DEZA, M. M.; DEZA, E. *Encyclopedia of Distances*. [S.l.]: Springer Berlin Heidelberg, 2009.

FACUNDES, V. L. D.; LUDERMIR, A. B. Transtornos mentais comuns em estudantes da área de saúde. *SciELO Analytics*, 2005.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. *AI Magazine*, v. 17, n. 3, p. 37, Mar. 1996. Disponível em: <https://www.aaai.org/ojs/index.php/aimagazine/article/view/1230>.

FREIRE, V. T.; OMAR, N. Comparação entre métodos, metodologias e frameworks para construção de sistemas computacionais analíticos-cognitivos. *Brazilian Journal of Development*, p. 31887–31904, 2020.

GIGLIO, J. S. Bem estar emocional em estudantes universitarios : um estudo preliminar. 1976.

GONCALVES, L. P. F.; FREITAS, H. Ferramentas de mineração de dados: Resultados efetivos? 2002.

GOWER, J. C. A general coefficient of similarity and some of its properties. *Biometrics*, [Wiley, International Biometric Society], v. 27, n. 4, p. 857–871, 1971. ISSN 0006341X, 15410420. Disponível em: <http://www.jstor.org/stable/2528823>.

HAN, J.; KAMBER, M.; PEI, J. *Data Mining*: Concepts and techniques. 3. ed. [S.l.]: Morgan Kaufmann, 2011.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of Statistical Learning:*: Data mining, inference, and prediction. [S.l.]: Springer, 2009.

HUANG, Z. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, p. 283304, 1998.

LOBO, M. B. d. C. et al. A evasão no ensino superior brasileiro. Instituto Lobo para o Desenvolvimento da Educação, da Ciência e da Tecnologia, 2007.

MAIMON, O.; ROKACH, L. *DATA MINING AND KNOWLEDGE DISCOVERY HANDBOOK*. 2. ed. [S.l.]: Springer, 2010. ISBN 978-0-387-09822-7.

MAJEED, I.; NAAZ, S. Current state of art of academic data mining and future vision. *Indian Journal of Computer Science and Engineering*, 2018.

MORAIS, E. A. M.; AMBRóSIO, A. P. L. Mineração de textos. 2007.

NASCIMENTO, R. L. S. do; JUNIOR, G. G. da C.; FAGUNDES, R. A. de A. Mineração de dados educacionais: Um estudo sobre indicadores da educação em bases de dados do inep. *Novas Tecnologias na Educação*, v. 16, n. 1, 2018. ISSN 1679-1916.

NEVES, M. C. C.; DALGALARRONDO, P. Transtornos mentais auto-referidos em estudantes universitários. *SciELO Analytics*, n. 4, p. 237–244, 2007.

NOGUEIRA, M. J. C. Saúde mental em estudantes do ensino superior: Fatores protetores e fatores de vulnerabilidade. 2017.

PELAEZ, K. et al. Using a latent class forest to identify at- risk students in higher education. *Journal of Educational Data Mining*, v. 11, n. 1, p. 18–46, 2019.

PRADO, H. et al. Predicting evasion candidates in higher education institutions. In: . [S.l.: s.n.], 2011. v. 6918, p. 143–151.

RAMOS, J. L. C. et al. Crisp-edm: uma proposta de adaptação do modelo crisp-dm para mineração de dados educacionais. *Sociedade Brasileira de Computação*, p. 1092–1101, 2020.

REYNOLDS, A. P. et al. Clustering rules: A comparison of partitioning and hierarchical clustering algorithms. *J. Math. Model. Algorithms*, v. 5, n. 4, p. 475–504, 2006. Disponível em: <http://dblp.uni-trier.de/db/journals/jmma/jmma5.htmlReynoldsRIR06>.

SHENG, W.; LIU, X. A hybrid algorithm for k-medoid clustering of large data sets. In: *Proceedings of the 2004 Congress on Evolutionary Computation (IEEE Cat. No.04TH8753)*. [S.l.: s.n.], 2004. v. 1, p. 77–82 Vol.1.

SHETH, J.; PATEL, B. Best practices for adaptation of data mining techniques in education sector. *tional Journal of System and Information Technology,*, v. 3, n. 2, p. 186, 2010. ISSN 0974-3308.

SILVA, H. F. D.; MARQUES, W. Evasão na educação superior no brasil: desafio à gestão acadêmica. 2017.

SILVA, L. A. da; PERES, S. M.; BOSCARIOLI, C. *Introdução a Mineração de Dados*: Com aplicações em r. 1. ed. [S.l.]: Elsevier, 2016. ISBN 978-85-352-8446-1.

SOKOLOVA, M.; LAPALME, G. A systematic analysis of performance measures for classification tasks. *Information Processing Management*, v. 45, p. 427–437, 07 2009.

THURAISINGHAM, B. *Data Mining*: Technologies, techniques, tools, and trends. [S.l.]: CRC Press, 1999. ISBN 0-8493-1815-7.

TONTINI, G.; WALTER, S. A. Pode-se identificar a propensão e reduzir a evasão de alunos?: ações estratégicas e resultados táticos para instituições de ensino superior. *SciELO Analytics*, 2014.

UNIFEI. *NORMA 2.3.03 - NORMA DO PROGRAMA DE ASSISTÊNCIA ESTUDANTIL DA UNIFEI*. 2020. Disponível em: <https://www.unifei.edu.br/institucional/documentos-/normas>.

WEISS, S.; INDURKHYA, N. *Predictive Data Mining*. 1. ed. [S.l.]: Morgan Kaufmann, 1997. ISBN 9781558604032.

ZHANG, T.; RAMAKRISHNAN, R.; LIVNY, M. *BIRCH: An Efficient Data Clustering Method for Very Large Databases*. 1996.