UNIVERSIDADE FEDERAL DE OURO PRETO

Relatório Técnico

Aplicação de técnicas de mineração de dados para obtenção de conhecimento útil relacionado à problemática da alta taxa de evasão no ICEA.

Lineker Aguiar Alcântara, Janniele Aparecida Soares Araujo, Helen de Cassia Sousa da Costa Lima.

Relatório técnico submetido como requisito para o encerramento do projeto pró-ativa.

Setembro de 2021

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

IES Instituições de Ensino Superior

GPL *General Public License*

LIWC *Linguistic Inquiry and Word Count*

ICEA Instituto de Ciências Exatas e Aplicadas

API *Application Programming Interface*

PCA *Principal Component Analysis*

JSON *JavaScript Object Notation*

UFOP Universidade Federal de Ouro Preto

NLTK *Natural Language Toolkit*

KDD *Knowledge Discovery in Database*

WCSS *within-clusters sum-of-squares*

Capítulo 1

Introdução

O fenômeno da evasão nas Instituições de Ensino Superior (IES) não é um problema recente, mas ganhou projeção e passou a ser um fator de extrema visibilidade a partir da década de 1990 em razão do processo de expansão da educação superior ocorrido nas últimas três décadas (JUNIOR et al., 2019). Já para DIAS et al. (2010), evasão é um dos "males" que afligem as instituições de ensino e comenta em sua pesquisa que os índices de evasão nas instituições são muito altos e que vem sendo uma realidade cada vez mais presentes nas IES.

A problemática da evasão universitária é vista como um dos fatores mais preocupantes nas IES, e de modo geral, o Instituto de Ciências Exatas e Aplicadas (ICEA) vem enfrentando este problema. Segundo dados passados pela Seção de Ensino do ICEA no período de 2012 à 2018 o número total de discentes ingressados nos quatro cursos oferecidos no campus é de 3990 e houveram 1772 evasões, representando $44,41\%$. Foram diplomados 863 em um ou mais cursos na instituição, equivalentes a $32,31\%$. A média de evasão entre os anos medidos é de 199 alunos.

No que diz respeito ao impacto no futuro da sociedade como um todo, de Carvalho Melo Lobo (2012) afirma que "O abandono do aluno sem a finalização dos seus estudos representa uma perda social, de recursos e de tempo de todos os envolvidos no processo de ensino, pois perdeu aluno, seus professores, a instituição de ensino, o sistema de educação e toda a sociedade (ou seja, o País)".

Diante do seguinte cenário, compreender os motivos que proporcionam a evasão no campus, se mostra uma tarefa de grande necessidade e relevância.

1.1 Objetivos

O projeto em questão tem como objetivo geral a extração de conhecimento útil sobre a problemática da alta taxa de evasão presente no campus, a partir de aplicações de algoritmos de mineração de dados, aprendizagem de máquina e análises estatísticas.

Para o alcance do objetivo geral, objetivos específicos foram traçados de forma a facilitar a chegada até os resultados desejados, são eles:

- Obtenção de dados relativos aos alunos que evadiram da universidade.
- Realização de tratamentos e limpeza para utilização de técnicas de mineração de dados.
- Detecção das possíveis causas da evasão, e suas relações.
- Identificação de características semelhantes entre os grupos das possíveis causas da evasão.

1.2 Planejamento

Metodologia aplicada para o desenvolvimento das atividades nos meses de vigência do projeto.

- Fevereiro:
  - Realizar a revisão bibliográfica.
  - Definir um dicionário de dados com as possíveis variáveis a serem inseridos nos modelos de mineração de dados.
- Março:
  - Realizar o processamento e limpeza dos dados.
- Abril:
  - Identificar as técnicas de Mineração de Dados que melhor se aplicam ao problema.

  - Identificar os coeficientes de correlação das variáveis (possíveis variáveis tais como: período de evasão, idade, problemas com disciplinas, oportunidades externas, etc).

- Maio, Junho:
  - Ajustar e aplicar a modelagem da análise de sentimentos.
  - Aplicar as técnicas de mineração de dados.
- Julho:
  - Analisar os resultados para extrair conhecimento útil.
- Agosto e Setembro:
  - Escrever e apresentar os resultados.

Capítulo 2

Referencial Teórico

Neste capítulo serão abordados conceitos teóricos e referências que regeram o desenvolvimento do trabalho em questão. Primeiramente, abordaremos as bibliografias base de estudo que foram utilizadas, e que serviram de inspiração. Posteriormente, será feito uma contextualização dos métodos e tecnologias aplicadas, desde o processamento de dados até a análise dos resultados. Por último, uma análise descritiva dos algoritmos usados, de como e por quais motivos foram escolhidos, bem como seus resultados de saída.

2.1 Revisão Bibliográfica

No decorrer do ciclo acadêmico, tem-se por conhecimento empírico que os alunos de gradução passam constantemente por inúmeras alternâncias do seu estado emocional. Diversos motivos podem levar a esta consequência, desde pressões externas até o seu rendimento pessoal no curso. Dentre os fatores psicológicos que podem influenciar tais alternâncias, o trabalho de Alves Pinto (2019) visa relacionar padrões de ansiendade detectados através das respostas a um questionário aplicado para alunos dos cursos de Engenharia de Computação e Sistemas de Informação do ICEA juntamente com seus dados obtidos de seus respectivos perfis no *Instagram*.

Para execução e obtenção dos resultados, os autores aplicaram a técnica de *Principal Component Analysis* (PCA) para identificação dos fatores que representam a maior variabilidade dos dados obtidos da aplicação do questionário desenvolvido. Posteriormente, com o resultado do PCA, executaram o algoritmo de aprendizagem de máquina não

supervisionada *K-means*, a fins de indentificação de grupos semelhantes formados pelos alunos, e suas principais características. A obtenção dos dados do *Instagram* foi realizada através de um *crwaler* desenvolvido em Python, com os dados obtidos, aplicou-se a técnica de análise de sentimentos nos textos retornados através do software *SentiStrength*. Os resultados obtidos das etapas de análise de sentimentos e mineração de dados do questionário foram então cruzados para identificação de possíveis correlações e semelhanças.

O trabalho relacionado teve contribuição para o desenvolvimento e metodologia do projeto. As técnincas abordadas para análise de dados, como o PCA, sua funcionalidade e interpretação dos resultados, bem como a clusterização com o algoritmo *K-means*. No tocante à análise de sentimentos, foi importante observar o comportamento do *software SentiStrength* no seu modo de classificação binário, e como este poderia ser modificado a depender do contexto.

Em Zacarias (2019) os autores sugerem que com a imensa quantidade de informações sendo geradas atualmente dia após dia, a obtenção de conhecimento útil através desses dados tem-se tornado um produto altamente valioso para grandes empresas que visam ofertar serviços personalizados para seus clientes. Partindo deste propósito, os autores viram desenvolver um sistema de geração automática de perfis de usuários através de seus dados coletados na suas respectivas redes sociais.

A realização da coleta dos dados foi feita através de uma abordagem *Cross Domain*, com objetivo de captação de informações de fontes distintas, sendo essas *Twitter* e *Instagram*. A coleta dos dados foi executada utlizando uma *Application Programming Interface* (API) fornecida por cada uma das redes sociais. Os dados foram retornados em um *JavaScript Object Notation* (JSON) e posteriormente tratados através do módulo externo do Python, o *Natural Language Toolkit* (NLTK), combinado com módulos nativos para efetuar uma limpeza inicial dos textos, como remoção de espaços em branco e caracteres especiais; segmentação do texto para obtenção de termos unitários (*tokenização*); remoção das palavras irrelevantes (*stop words*) e obtenção do radical da palavra (*steaming*).

Após a pré processamento dos dados, foi feita a medição dos pesos de cada *token* através de medidas estatísticas KF-IDF (Xu et al., 2002) e o TF-IDF (Salton, 1991), correlacionando-os com o dicionário *Linguistic Inquiry and Word Count* (LIWC). Para inferência da personalidade, os *tokens* foram mapeados do LIWC para o modelo *Big Five* de personalidades, através da correlação entre ambos proposta por Schwartz et al.

(2013).Após a obtenção das personalidades, foi aplicado o algoritmo de aprendizagem de máquina, *KNN*, para $K = 3$, com objetivo de realizar classificações de personalidades para novas entradas aplicadas ao modelo, alcançando um taxa de acerto de $58,6\%$ como resultado final.

O trabalho analisado seviu como base no que se refere a análise de sentimentos e pré processamentos em dados não estruturados. Podemos mencionar também a realização dos processamentos de análise de frequência dos *tokens* em diferentes métricas estatísticas.

de Oliveira (2021) sugere que com vasto número de usuários nas redes sociais atualmente e o estreitamento de suas interações permitem um grande compartilhamento de ideias e opiniões sobre os mais diversos assuntos. As eleições presidenciais brasileiras, ocorridas em 2018 se mostrou também como uma oportunidade para realização de estudos sobre o comportamento dos usuários brasileiros nas redes sociais. O trabalho desenvolvido pelo autor teve como objetivo analisar os sentimentos das postagens e comentários no *Facebook* relacionados aos candidatos Jair Bolsonaro e Fernando Haddad, antes, duarante e após as eleições, a fim de detectar polarizações e padrões de sentimentos relacionados aos candidatos.

Os autores coletaram os dados através do *Graph API* [1] fornecida pela própria plataforma do *Facebook* para desenvolvedores. Após a obtenção, aplicou-se algoritmos de limpeza de dados em conjunto com as funções oferecidas pelo módulo NLTK para processamento de linguagem natural. Devidamente processados, os dados foram submetidos ao *software* de análise de sentimentos, o *SentiStrength*. Os resultados de saída foram então passados para uma análise mais profunda para maior entendimento dos sentimentos relacionados com os candidatos em cada um dos períodos da eleição.

Diversas técnicas utilizadas pelos autores é de grande importância para a fundamentação prática deste projeto, como por exemplo a maior compreensão sobre os diversos métodos de análise de sentimentos existentes e suas especificidades, juntamente com aplicação de algoritmos para limpeza de dados textuais. Observou-se também o comportamento do *SentiStrength* para análises em português, bem como seus diversos modos de classificação, além do impacto positivo causado pela alteração do vocabulário base do *software*.

[1] `https://developers.facebook.com/docs/graph-api/`

2.2 Ferramentas e Métodos

Dada a grande quantidade de ferramentas existentes no cenário atual para desenvolvimento de projetos em diferentes âmbitos, a escolha dessas ferramentas devem convergir para as necessidades de cada trabalho, de forma a serem melhor utilizadas e possibilitar uma maior performance. Assim sendo, as tecnologias descritas são de fundamental importância não só para alcançar os objetivos, mas também obter versatilidade e aceleração do projeto, facilitando na aplicação dos métodos nos diferentes estágios que serão abordados a seguir.

2.2.1 Linguagem de programação Python

O Python possui uma sintaxe clara e concisa que fornece a legibilidade do código-fonte, tornando a linguangem mais produtiva. A mesma, inclui diversas estruturas de alto nível, uma vasta coleção de módulos internos prontos para uso, além de *framworks* de terceiros que podem ser adicionados (Borges, 2014).

Os benefícios considerados do uso do python para a execução do projeto, vão além da facilidade para escrita e leitura dos *scripts*. Atualmente, a linguagem está em código aberto (com a licença compatível com a *General Public License* (GPL), porém menos restritiva, permitindo que seja inclusive incorporada em produtos proprietários), multiplataforma e interpretada, sendo capaz de ser executada em qualquer sistema operacional e multiparadigma, suportando programação modular, funcional e orientada a objetos (Borges, 2014).

Com o vasto uso na área científica, existem atualmente diversos ambientes interativos compartilhados que permitem a execução de código python em servidores remotos, com inúmeros módulos externos pré-instalados (com propósitos relacionados a ciência de dados e afins). Dentre tais módulos, podemos citar aqui alguns essenciais para o desenvolvimento e conclusão do projeto, são eles: Pandas[2], para manipulação e análise dos dados; *Numpy*[3], para realização de cálculos matemáticos; *Scikit-learn*[4], biblioteca que oference inúmeros algoritmos de aprendizagem de máquina e tratamento de dados; NLTK[5]; utilizada para tratamentos de textos em linguagem natural; *Matplotlib*[6],

[2] `https://pandas.pydata.org/`
[3] `https://numpy.org/`
[4] `https://scikit-learn.org/stable/`
[5] `https://www.nltk.org/`
[6] `https://matplotlib.org/`

Seaborn[7] e *Plotly*[8] para vizualização de dados.

As vantagens trazidas pelas características intrínsecas à linguagem aqui relatada, bem como todos os seus módulos externos disponíveis para ciência de dados são de suma importância para a realização do projeto.

2.2.2 SentiStrenth

O *SentiStrength*[9] é um software para análise de sentimentos que utiliza como base original um dicionário léxico de 2310 palavras e radicais obtidos do *Linguistic Inquiry and Word Count* (LIWC) (Thelwall, 2013). Tal dicionário léxico tendo os sentimentos de cada termo previamente classificados por humanos, e posteriormente aprimorado com uso de aprendizagem de máquina.

Segundo Thelwall (2013), o software aqui utilizado possui diversas possíveis saídas como retorno da classificação, sendo elas:

- Binário: positivo ou negativo, sendo associados aos valores 1 e -1 respectivamente.
- Trinário: objeto com três entradas, (positivo/negativo/neutro), sendo do mesmo funcionamento do modo binário, com a adição do valor de neutralidade.
- Em escala: único valor variando no intervalo de $[-4, 4]$.
- Dual: retornando um objeto com dois valores, o primeiro relacionado ao sentimento positivo no intervalo de $[1, 5]$, e o último ao sentimento negativo, no intervalo de $[-1, -5]$. Como ilustrado na Tabela 2.1

Além da base de dados léxica originalmente desenvolvida para o indioma inglês, o *SentStrength* também conta com outros dicionários disponíveis para análises em diferentes linguas, contudo, podendo haver uma queda nos resultados finais. Segundo Reis et al. (2016) o *software* possui uma taxa de 82% de performance para identificação de sentimentos em abordagens multilíngues para o português. Importante ressaltar que tal

[7] `https://seaborn.pydata.org/`
[8] `https://plotly.com/`
[9] `http://sentistrength.wlv.ac.uk/`

Texto	Avaliação termo a termo	Valores retornados
Eu amo minha família	Eu[0] amo[2] minha[0] familia[0]	$[3, -1]$
Eu não amo minha família	Eu[0] não [0] amo[2] minha[0] familia[0]	$[1, -3]$
Eu odeio chocolate	Eu[0] odeio[-3] chocolate[0]	$[1, -4]$

Tabela 2.1: Classificação do *SentiStrength* no modo dual

performance pode ser melhorada com a modificação das base de dados em qualquer indioma, dado a especificidade de cada projeto, processo realizado diante do nosso contexto e será explicado posteriormente.

Dentre os arquivos que o *software* utiliza para realizar a análise, podemos citar: *EmotionLookupTable.txt* que contém as relações de todas as palavras ali presentes com seu valor para classificação, caso verificada sua ocorrência na frase. *BoosterWordList* que contém os termos que podem intensificar os sentimentos, como por exemplo o termo "muito". *NegatingWordList* que contém as palavras de negação que auxiliam para inversão de sentimentos, como exemplificado na Tabela 2.1; e, arquivos para detecção de ironia, gírias e emojis.

2.2.3 Googletrans

O Googletrans[10] é uma biblioteca python gratuita e ilimitada que implementa a *Application Programming Interface* (API) do *Google Translate*. A biblioteca utiliza *Google Translate Ajax API*[11] para fazer chamadas para métodos como detectar o indioma e traduzir. A máquina de tradução utilizada possui confiabilidade já verificada por outros trabalhos (Balahur e Turchi, 2012).

[10] `https://pypi.org/project/googletrans/`
[11] `https://translate.google.com/`

2.3 Métodos Aplicados

2.3.1 Análise de Sentimentos

O principal foco ao realizar um processo de análise de sentimentos está na automatização de extração de informações subjetivas de textos em linguagem natural, como oponiões e sentimentos (Benevenuto et al., 2015). Tais extrações, para Liu e Zhang (2012) se da pelo estudo computacional das opiniões, avaliações, atitudes e emoções descritas por pessoas na *web*, relacionadas a diferentes tópicos. A análise de sentimento, ou do inglês *sentiment analysis*, vem crescendo continuamente desde o início da expanção das redes sociais no fim dos anos 2000 (Benevenuto et al., 2015). A Figura 2.1 ilustra esse aumento do interesse ao longo do tempo em todo o mundo até os dias atuais.

Figura 2.1: Interesse ao longo do tempo do termo *Sentiment Analysis* no Google Trends

Para Balahur e Turchi (2012), existem atualmente três principais abordagens para realização de uma classificação de sentimentos: técnicas não supervisionadas, que utilizam dicionários léxicos com medidas quantitativas de sentimentos positivos e negativos associados às palavras, além de um conjunto de regras para calcular o resultado final da sentença; técnicas supervisionadas, que se baseiam na criação de modelos de aprendizagem de máquina sob um conjunto palavras de treino previamente classificadas; e, as técnicas semi supervisionadas, nas quais empregam ambas as estratégias ateriormente citadas, utilizam dicionários léxicos com valores previamente definidos para posteriormente realizar o emprego de modelos de aprendizagem de máquina para aprimoramento da classificação desses exemplos de treino.

A abordagem supervisionada exige uma etapa de treinamento de um modelo com amostras previamente classificadas, sendo este (a obtenção de dados rotulados para treino e teste) um dos estágios para seu funcionamento, além da definição das *features*

para distinção dos dados, treinamento do modelo de aprendizagem de máquina, e como último estágio, sua aplicação (Benevenuto et al., 2015). Para Balahur e Turchi (2012), a etapa de obtenção de dados rotulados confiáveis pode ser o maior desafio para aplicação desta estratégia.

Para a execução das técnicas não supervisionadas, o maior contraponto pode estar na obtenção de dicionários léxicos grandes o sufucientes para lidar com a variabilidade da língua, que pode ser muito caro se feito manualmente, ou não confiável, se feito automaticamente (Balahur e Turchi, 2012). Contudo, esta técnica não necessita de sentenças previamente rotuladas, dessa forma, não se restrigindo ao contexto na qual foi empregada (Reis et al., 2016). Podemos analisar o funcionamento desta estratégia, resumidamente, de acordo a Figura 2.2.

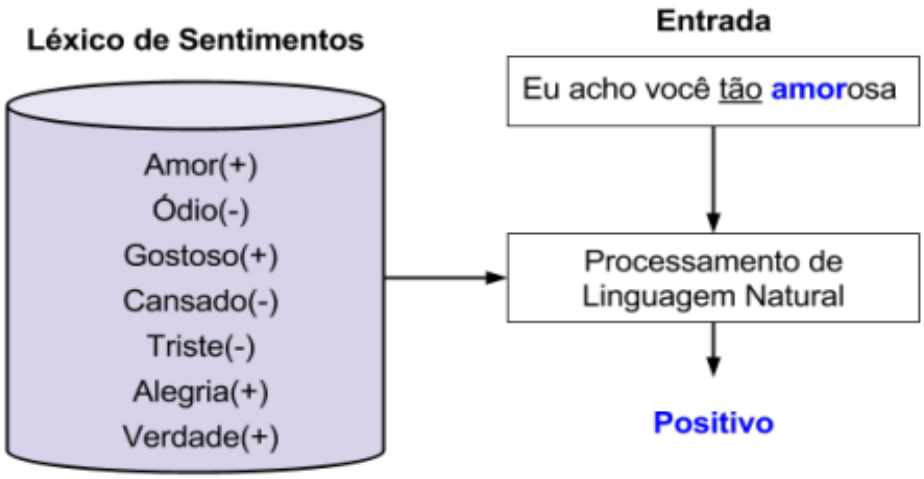


Figura 2.2: Análise léxica para os métodos não supervisionados
Fonte: (Benevenuto et al., 2015)

Já as abordagens semi supervisionadas, dependem altamente do desempenho do conjunto inicial de exemplos classificados. Se tais exemplos possuírem erros na rotulação, gerados por exemplo por traduções, esses erros neste pequeno conjunto inicial teria um alto impacto negativo no resultado final dos exemplos subsequentes rotulados (Balahur e Turchi, 2012).

2.3.2 Knowledge Discovery in Database (KDD)

O desenvolvimento do trabalho foi realizado a partir dos conceitos abordados do processo de descoberta de conhecimento em base de dados, do inglês, *Knowledge Discovery in Database* (KDD). Segundo Fayyad et al. (1996), o KDD se baseia no desenvolvimento e aplicação de métodos para dar sentido aos dados analisados, sendo mapeados de seu estado bruto para formas mais descritivias e de fácil análise. Ainda que no cerne da técninca a esteja a aplicação de técnicas de mineração de dados para descoberta e extração de padrões, as etapas de preparação, seleção e limpeza dos dados, incorporação de conhecimento e interpretação adequada dos resultados são essenciais para garantia de que conhecimentos úteis serão extraídos.

As etapas básicas do KDD são ilustradas pela Figura 2.3, em que Fayyad et al. (1996) as descreve de maneira que o primeiro passo se da na seleção do conjunto de dados como fono no subconjunto de variáveis em que o processo será aplicado; pré-processamento: execução de estratégias para limpeza de dados ruidosos ou ausentes, entre outros; tranformação: procura por atributos úteis visando o objetivo da pesquisa, aplicação de métodos para redução do número de variáveis ou representação invariante dos dados; mineração de dados: aplicação de algoritmos específicos para extração de padrões dos dados; e, avaliação: interpretação dos padrões econtrados, podendo ser necessário refazer todas as etapas anteriores. Ainda segundo o autor, os algoritmos de mineração de dados podem ser classificados da seguinte forma:

- Classificação: aprendizagem de uma função que mapeie os dados em uma ou várias classes pré-definidas.
- Regressão: aprendizagem de uma função que mapeie os dados em valores pertecentes ao conjunto dos número reais.
- Clusterização: busca a identificação de um conjunto finito de grupos ou categorias para descrever os dados.
- Sumarização: aplicação de métodos para obter uma descrição de um subconjunto dos dados.
- Modelagem de dependências: encontrar modelos que descrevam dependências significantes entre as variáveis.
- Detecção de alterações ou desvios: descoberta de mudanças mais significativas nos dados a partir de valores previamente medidos.

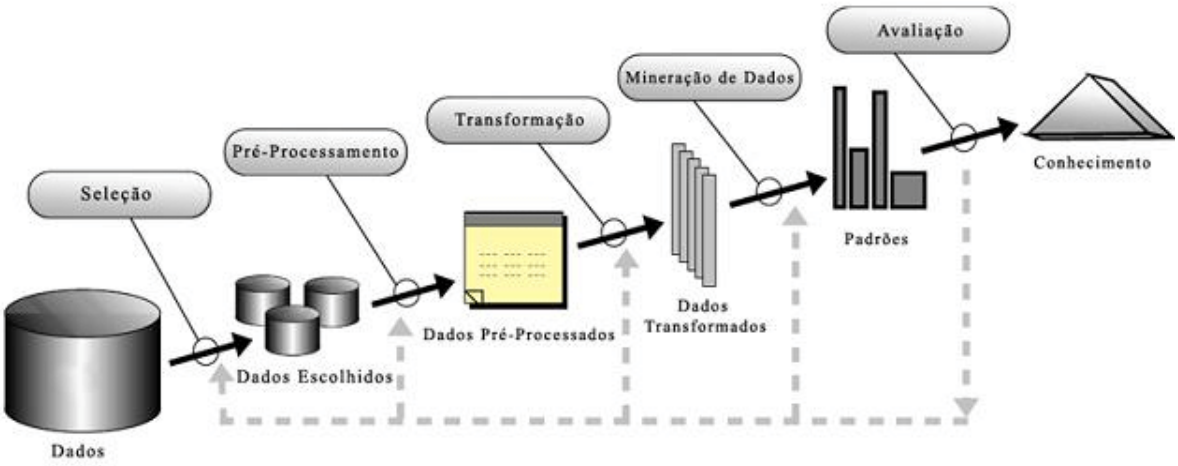


Figura 2.3: Etapas do KDD
Fonte: (Fayyad et al., 1996)

2.3.3 Medidas de Correlação

A correlação é a medida de uma possível associação linear entre duas variáveis. Podemos ter uma correlação positiva, onde observaremos um aumento ou diminuição no valor de uma variável quando a outra aumentar ou diminuir, respectivamente; uma correlação negativa, quando o valor de uma variável aumenta, a outra diminui; e uma correlação fraca, de forma que as mudanças de valores de uma variável não influenciam a outra. (Schober et al., 2018).

Segundo Mukaka (2012), a correlação é uma medição adimensional que assume valores no intervalo de $[-1, 1]$. Um coeficiente de correlação em zero indica que não existe nenhuma relação linear entre as duas variáveis, já para os valores de -1 ou 1 indica uma correlação linear perfeita. A força dessa associação linear será observada através dos coeficientes retornados, quanto mais próximos os coeficentes estiverem dos valores limites de ± 1, mais fortes serão as correlações.

Se os coeficientes forem positivos, as variáveis serão diretamente relacionadas, ou seja, um aumento ou diminuição da magnitude em uma variável implicará no aumento ou diminuição da outra variável correspondente, respectivamente. Para coeficientes negativos, teremos variáveis inversamente relacionadas, o aumento no valor de uma variável causará o decréscimo da outra, bem como o caminho inverso, onde o decréscimo de uma variável implicará no aumento da outra relacionada.

Existem atualmente dois métodos principais para o cálculo de correlação: Pearson e Spearman (Mukaka, 2012). A correlação de Pearson requer que as variáveis analisadas sejam derivadas de uma amostra aleatória ou de uma amostra representativa dos dados, além de serem contínuas e normalmente distribuídas (Schober et al., 2018). A correlação de Spearman não requer que as variáveis sejam contínuas, e não se restringe à análises de variavéis normalmente distribuídas, podendo ser utilizada também em dados ordinais (Schober et al., 2018).

2.3.4 Principal Component Analysis (PCA)

Conjuntos de dados de alta dimensionalidade se tornaram cada vez mais comuns e muitas vezes difíceis de serem interpretados. A aplicação do PCA surge como uma técnica para uma redução de dimensionalidade de tais conjuntos, gerando uma nova formação de variáveis de forma a aumentar o poder de interpretação e ao mesmo tempo minimizando a perda de informação causada pela redução de dimensionalidade (Jolliffe e Cadima, 2016).

As dificuldades na interpretação de dados de muitas dimensões podem passar desde a visualização gráfica para o comportamento das variáveis e como eles se ditribuem, até na retirada de medidas estatíticas de interesse. Além de tais problemas citados, existem também obstáculos para a aplicações de modelos de aprendizagem de máquina, que podem ser causados, por exemplo, devido a multicolinearidade de dados, onde um modelo pode incluir variáveis redundantes para o algoritmo.

O PCA é uma técnica multivariada que analisa os dados de obsevações que são descritas por diversas variáveis dependentes, quantitativas e correlacionadas. O objetivo principal é representar essas informações em um novo conjunto de novas variáveis ortogonais denominadas componentes principais (Abdi e Williams, 2010).

Ao realizar a aplicação da técnica aqui descrita, poderá ser obtida uma nova matriz com as componentes principais organizadas em ordem decrescente, de acordo variabilidade dos dados explicada por cada uma delas. Dessa forma, pode ser escolhido a quantidade de componentes principais necessárias parar explicar o valor desejado da variabilidade total dos dados. Além da nova matriz, podemos obter informações importantes como os *outliers* e também as variáveis de maior explicabilidade dos dados originais.

2.3.5 Clusterização

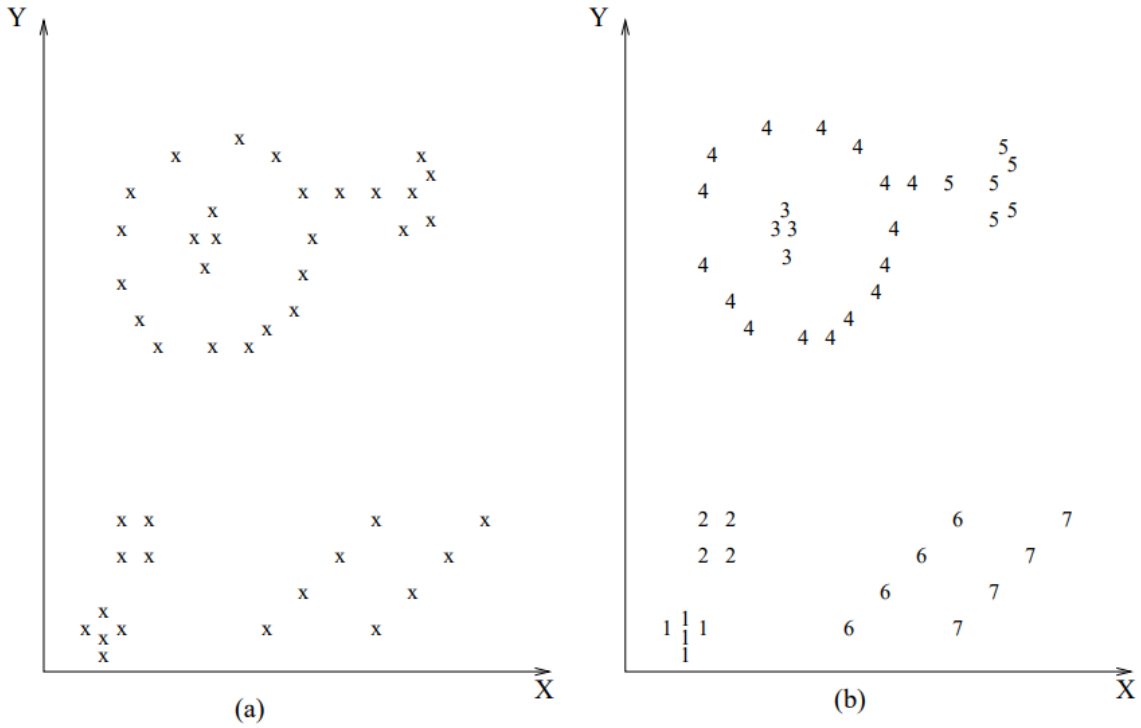
De acordo Ochi et al. (2004), a clusterização, ou agrupamento, se baseia no método de alocar objetos de uma dada base de dados, de modo que os objetos mais similares sejam agrupados em um mesmo cluster, e objetos menos similares sejam alocados em clusters distintos. Em uma definição mais concisa, Doni (2004), explica que objetivo do algoritmo é que os clusters finais sejam determinados de forma que haja uma homogeneidade entre seus elementos, e uma heterogeneidade entre os grupos.

A clusterização é largamente aplicada em diferentes campos com diversos propósitos finais, podemos utilizar a aplicaçao da técnica em problemas envolvendo análise e identificação de padrões; agrupamento; tomadas de decisão; situações envolvendo apredizagem de máquina, como mineração de dados e segmentação de imagens; além de classificação de padrões (Jain et al., 1999). Segundo Ochi et al. (2004), existem atualmente duas classes relacionadas a tal abordagem: o *Problema de K - Clusterização*, classe mais estudada e aplicada, que é onde o número dos K clusters já são previamente defenidos, e o *Problema da Clusterização Automática*, onde não é predefinido o número dos K clusters desejados, de tal forma que o próprio algoritmo seja capaz de identificar a quantidade ideal de grupos.

Para realizar uma definição mais formal, Ochi et al. (2004) explica que dado um conjunto de dados X com n observações, $X = \{X_1, X_2, X_3, ..., X_n\}$, o objetivo do método é a formação do conjunto de *clusters* C com k *clusters*, $C = \{C_1, C_2, C_3, ..., C_k\}$. Contudo, os subconjuntos de C precisam satisfazer três propriedades: não podem ser vazios; precisam ser disjuntos entre si e a operaçao de união aplicada em C, deve retornar o conjunto original X. Podemos observar a Figura 2.4 como exemplo de uma aplicação genérica do algoritmo, são obtidos sete *clusters*, cada um sendo representado pelo seu respectivo número.

A realização das medidas de similaridade entre objetos, ou observações, possuem inúmeras abordagens e métricas. A definição de qual métrica utilizar ao empregar o algoritmo pode variar de acordo os objetivos e necessidades de cada trabalho.

Figura 2.4: Clsterização de dados



Fonte: (Jain et al., 1999)

Capítulo 3

Desenvolvimento

Descreveremos aqui todos passos realizados para o desenvolvimento do trabalho seguindo a metodologia KDD, descrita no Capítulo 2. Primeiramente será explanado as etapas de levantamento, seleção e pré-processamento dos dados, porteriormente, as etapas de transformação e mineração de dados e por último a análise dos padrões encontrados.

3.1 Levantamento e pré-processamento de dados

Para levantamento dos dados relacionados ao problema, foi elaborado um questionário composto pelas seguintes perguntas:

1. Informe o ano que você evadiu (saiu) da UFOP).
2. Período de evasão.
3. Sexo.
4. Idade (idade em que saiu do curso).
5. Qual período você estava quando saiu?
6. Selecione o(s) possíveis motivos que levou você a abandonar o curso:
    - Problemas financeiros.
    - Problemas familiares.

- Problemas psicológicos.
- Perda(s) de entes queridos.
- Fim de relacionamento.
- Distanciamento da família.
- Oportunidade de ingresso em outra faculdade com mesmo curso.
- Oportunidade de ingresso em outra faculdade em curso diferente.
- Oportunidade de ingresso no ICEA em curso diferente.
- Oportunidade de ingresso na UFOP em curso diferente.
- Problema de adaptação com o curso.
- Dificuldade na área escolhida.
- Desinteresse com curso escolhido.
- Desinteresse com curso superior em geral.
- Oportunidade de trabalho.
- Incompatibilidade de tempo entre trabalho e estudo.
- Faixa salarial não promissora.
- Baixo rendimento nas disciplinas.
- Problemas com uma disciplina em específico.
- Problemas com professores específicos.
- Desligamento por prazo máximo.
- Desligamento por reprovações em todas as disciplinas dois semestres consecutivos.
- Desligamento por coeficiente abaixo de 3 dois semestres consecutivos.
- Cenário político/enconômico instável.
- Problemas com aprendizado em relação ao conteúdo.

7. Descreva uma frase que levou você a largar o curso.

O questionário foi elaborado através da plataforma do *Google Forms*[1] e submetido para os alunos que já evadiram de todos os cursos oferecidos no campus do ICEA através

[1] `https://www.google.com/intl/pt-BR/forms/about/`

de seus *e-mails* institucionais, obtendo um total de 167 respostas. A base de dados formada através do questionário possui uma estrutra como a ilustrada pela Tabela 3.1. Na seleção dos atributos, devido ao caráter da pesquisa, os atributos "ano de evsão" e "semestre de evasão" foram desconsiderados para a etapa de mineração de dados e avaliação.

Atributo	**Descrição**	**Exemplo**
Ano de evasão	Ano em que o aluno evadiu da UFOP	2015
Semestre de evasão	Semestre do ano em que o aluno evadiu da UFOP	1º ou 2º Semestre
Sexo	Sexo do aluno	Masculino, feminino ou outro
Idade	Idade que o aluno tinha ao evadir da universidade	23
Período de evasão	Período do curso que o aluno estava ao evadir da universidade	1 a 10, ou superior a 10
Motivos de evasão	Lista com 25 possíveis motivos que levou o aluno a evadir	Problemas financeiros; oportunidade de trabalho; baixo rendimento nas disciplinas
Justificativa de evasão	Justificativa com as próprias palavras do aluno pela evasão	Passei em outra universidade próxima da minha família

Tabela 3.1: Descrição da base de dados do questionário

Para o pré-processamento, inicialmente foi verificado e corrigido a existência de possíveis erros no conjunto de dados. Devido aos valores faltantes no atributo de texto livre de justificativa de evasão, e dados duplicados, o conjunto inicial com 167 observações foi reduzido para um total de 134 observações.

Após a inspeção dos erros, foi realizado o tratamento para o atributo contendo os 25 possíveis motivos de evasão, visando transformar cada um dos motivos em um novo atributo para a base de dados através da construção de uma matriz de incidência, foi utilizada a estratégia do *bag-of-words*. Cada motivo de evasão foi transformado em um só termo, como "problemas psicológicos" para "problemaspsicológicos", e utilizando o

método *CountVectorizer*[2], obtivemos uma matriz com os 25 novos atributos demarcados com 1 para os alunos que relataram o problema associado a nova variável, e 0 caso contrário.

A fim de lidar com dados em linguagem natural deixados pelos alunos no campo de texto livre. Primeiramente, foram retirados todos os caracteres não ASCII de cada frase e em seguida convertidas cada uma para caixa baixa. Atraves do módulo re[3], nativo do python para tratamento de expressões regulares, foram retirados os caracteres não pertecentes ao alfabeto da língua portuguesa. Por último, realizou-se a tokenização de cada frase através do método *word_tokenize* e removidas todas as palavras que não possuem relevância para o tratamento dos textos, as *stop words*, juntamete com as pontuações presentes.

Após a etapa de limpeza dos textos, utilizando novamente o método *CountVectorizer*, realizou-se a contagem da frequência dos n-gramas no texto. Segundo Moreira e Favero (2009), um n-grama consiste em uma sequência contígua de n caracteres extraídos de uma amostra de texto. Dessa forma, foi analisado a frequência para $n = 1$, como os unigramas, "estudo", "trabalho", e $n = 2$, para os bigramas como "outro curso", "outra universidade".

Com objetivo da extração de informações dos textos tratados, as sentenças foram submetidas a uma análise de sentimentos realizada através do *software SentiStrength*. Com intuito da realiazação de uma abordagem multilíngue, inicialmente, todas os textos em eu estado bruto foram traduzidos para a língua inglesa pela biblioteca gratuita do python, *Googletrans*. Posteriormente, para verificação de possíveis erros de tradução, todas as novas sentenças passaram por uma inspeção manual para efetuar as correções necessárias. Por último, foram aplicadas as mesmas etapas descritas de limpeza dos textos em português para os textos traduzidos, e então realizada a contagem das frequências dos unigramas e bigramas presentes.

Como observado em de Oliveira (2021), a detecção de sentimentos realizada pelo *SentiStrength* pode ser melhorada ao realizar ajustes em sua base de dados de acordo o vocabulário dos textos que está sendo analisado. Dessa forma, dada as frequências

[2] `https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html`

[3] `https://docs.python.org/3/library/re.html`

Sentido original	**Palavras relacionadas**	**Frequência**	**Valor associado**
Falta	falta, faltou, faltando	$19 + 2 + 1 = 22$	-3
Trabalho	trabalho, trabalha, trabalhando, trabalhar, trabalham, trabalhei, trabalhavam	$11 + 3 + 2 + 2 + 1 + 1 + 1 = 21$	-2
Outro	outra, outro, outros, outras	$9 + 5 + 2 = 17$	1
Dificuldade	dificuldade, dificuldades, dificulta, dificultam	$8+1+1+1 = 11$	-2
Problema	problemas, problema	$7 + 3 = 10$	-2
Família	família, pais, familiares	$5 + 2 + 2 = 9$	-1
Mudança	mudança, mudei, mudar	$5 + 2 + 1 = 8$	2
Passar	passei, passar	$5 + 2 = 7$	2
Greve	greve, greves	$4 + 3 = 7$	-3
Reprovação	reprovação, reprovam, reprovados, reprovar, pau	$2+1+1+1+1 = 6$	-3
Gostar	gostando, gostei	$3 + 2 = 5$	3
Adaptação	adaptei, adaptação	$3 + 2 = 5$	2
Coonseguir	consegui, conseguia	$3 + 2 = 5$	2
Aprovação	aprovação, aprovado	$4 + 1 = 5$	2
Distância	distância, longe	$3 + 1 = 4$	-2
Depressão	depressão	3	-4
Incompatibilidade	incompatibilidade	3	-3
Desinteresse	desinteresse	3	-2

Tabela 3.2: Relação das alterações na base de dados do *SentiStrength* no arquivo *EmotionLookupTable* para análise em português.

observadas na contagem dos unigramas, as bases de dados do *software* sofreram os ajustes de inserção de palavras e alteração do valor do sentimento associado, para a análise em português descritos pelas Tabelas 3.2, 3.3, e para a análise em inglês descritos pelas

Sentido original	**Palavras relacionadas**	**Frequência**	**Valor associado**
Baixo	baixo, baixa, baixas	$5+1+1=7$	-2
Pouco	pouco, pouca, poucas	$2+1+1=4$	-3
Péssimo	péssimos, péssimas	$1+2=3$	-4
Ruim	ruim, ruins	$1+1=2$	-3

Tabela 3.3: Relação das alterações na base de dados do *SentiStrength* no arquivo *BoosterWordList* para análise em português.

Sentido original	**Palavras relacionadas**	**Frequência**	**Valor associado**
Lack	Lack, lacked, lacking	$18+1+1=22$	-3
Work	Work, working, worked	$14+5+2=21$	-2
Another	another, other	$13+4=17$	1
Difficulty	difficulty, difficult, difficulties	$8+2+1=11$	-2
Problem	problems, problem	$8+2=10$	-3
Family	family, parents	$7+2=9$	-1
Change	change, moved, move	$5+2+1=8$	2
Strike	strike, strikes	$4+3=7$	-3
Fail	fail, failing	$3+3=6$	-3
Pass	passed, pass	$5+1=6$	2
Adapt	adapting, adaptation, fit, adapt	$1+1+1+2=5$	2
To get	get, getting, got	$3+1+1=5$	2
Approve	approval, approved	$4+1=5$	2
Distance	distance, away	$3+1=4$	-2
Incompatibility	incompatibility	3	-3
Depression	depression	3	-4

Tabela 3.4: Relação das alterações na base de dados do *SentiStrength* no arquivo *EmotionLookupTable* para análise em inglês.

Tabelas 3.4, 3.5. Após os ajustes, a análise de sentimentos foi executada sobre os textos e os sentimentos associados a cada observação foram adicionados ao conjuntos de dados,

Sentido original	**Palavras relacionadas**	**Frequência**	**Valor associado**
Low	low	6	-3
Little	little, few	$3+1$	-2
Bad	bad, terrible	$3+2=5$	-2, -3

Tabela 3.5: Relação das alterações na base de dados do *SentiStrength* no arquivo *BoosterWordList* para análise em português.

resultando na criação de quatro atributos, sentimento positivo e negativo relacionados aos textos em ambos idiomas.

3.2 Transformação e mineração de dados

Finalizado o pré-processamento, iniciou-se a fase de aplicação de métodos para transformação e mineração de dados, definido por Fayyad et al. (1996) como a busca por padrões e formas mais representacionais através de aplicações de algoritmos de agrupamento, classificação, sumarização e regressão e posteriormente a interpretação de tais padrões encontrados.

O primeiro processo para extração de conhecimento útil foi entender o relacionamento das variáveis remanescentes da base de dados, aplicando assim a análise de correlação através do método de spearman, descrito no Capítulo 2. Dentre todos os valores retornados, destacou-se as correlações positivas e negativas para os seguintes atributos:

- Correlações positivas:
  - Desligamento por reprovações em todas disciplinas dois semestres consecutivos e desligamento por coeficiente abaixo de 3 dois semestres consecutivos.

    A alta correlação observada entre os atributos pode ser facilmente explicada, pois de acordo a quantidade de reprovações em semestres consecutivos, fatalmente a chance de ser desligado por um baixo coeficiente aumentará devido as quedas nas notas.
  - Dificuldade na área escolhida e baixo rendimento nas disciplinas.

    Observamos aqui que, quanto maior a dificuldade na área escolhida, maior será os relatos de baixo rendimento nas disciplinas.

  - Problemas com aprendizado em relação ao conteúdo e baixo rendimento nas disciplinas.

    A correlação entre esses dois atributos nos mostra como os problemas com o aprendizado podem acarretar em um alto índice de baixo rendimento nas disciplinas.

  - Problemas de adaptação com o curso e dificuldade na área escolhida.

    Podemos notar que, de acordo ocorre um aumento nos problemas de adaptação com o curso, a dificuldade na área escolhida também aumentará.

  - Problemas com aprendizado em relação ao conteúdo e e dificuldade na área escolhida.

    Da mesma forma que problemas com o aprendizado podem ocasionar em uma maior dificuldade na área escolhida.

- Correlações negativas:

  - Para as correlações negativas, observamos um baixo coeficiente de correlação com a oportunidade de ingresso em outra faculdade com mesmo curso e os atributos de baixo rendimento nas disciplinas; desinteresse com curso escolhido; dificuldade na área escolhida e problema de adaptação com o curso.

    O que pode ser interpretado como, por exemplo, um aumento em quaisquer um dos atributos acarretará em uma diminuição dos valores de oportunidade de ingresso em outra faculdade, ou seja, o problema está associado diretamente com o curso, pois uma vez que o aluno tenha problemas, não buscará o mesmo curso em outra universidade.

  - Idade de evasão e oportunidade de ingresso em outra faculdade em um curso diferente.

    Analisando o coeficiente de correlação negativo entre os atributos, notemos que, ao observarmos uma dimuição nas idades dos alunos ao evadir do campus (alunos mais jovens), notaremos mais relatos de oportunidades em ingresso em outra faculdade em um curso diferente.

  - Problemas com professores específicos e sentimento negativo em inglês.

    Ressaltamos aqui que, o intervalo de variação do sentimento negativo pertence aos números negativos, ou seja, quanto menor o valor observado, mais distante do zero, mais negativo será o sentimento. Dessa forma, entendemos que, havendo grandes valores em problemas com professores específicos, menores serão os valores de sentimento negativo, ou seja, serão mais intensos.

– Período de evasão e sentimento negativo inglês.

  Seguindo o mesmo raciocínio, para maiores valores de período de evasão, observaremos sentimentos negativos mais intensos.

Devido a alta dimensionalidade da base de dados final, com o propóitos de reduzir a quantidade de variáveis e identificação das que mais representam a variabilidade dos dados, foi aplicada a técnica de análise de componentes principais, do inglês, *Principal Component Analysis* (PCA). Para escolha da quantidade de componentes que representarão os dados, foi utilzada a métrica do *scree plot*. A métrica consiste em plotar os autovalores em ordem decrescente e examinar onde ocorre uma quebra de nivelamento da linha tracejada, o número ideal é indicado através da quantidade de componentes anterior ao ponto de quebra (Kanyongo, 2005).

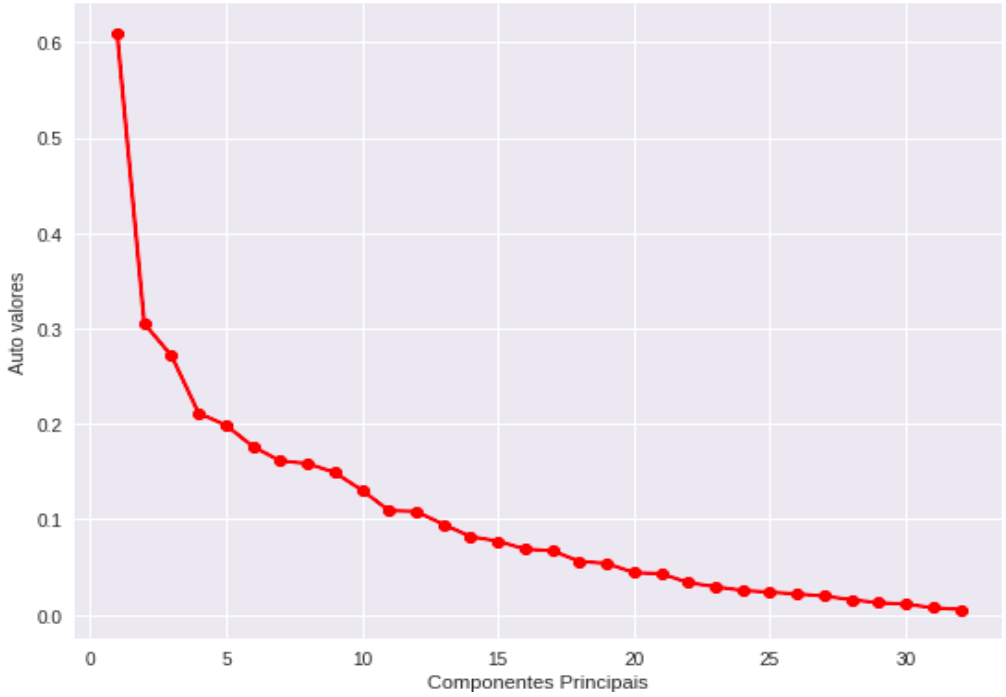


Figura 3.1: Scree Plot

Após a aplicação da técnica, foi obtido o gráfico ilustrado pela Figura 3.1. Devido a dificuldade em observar uma quebra clara da direção da linha traceada, o número de componentes foi escolhido de forma que mantivesse no mínimo 80% da variabilidade dos dados, assim sendo, foram necessárias 14 componentes principais preservando cerca de 81.7% das informações originais.

Visando a identificação de padrões formados pelas resposas dos alunos após a redução

de dimensionalidade, o conjunto de componentes resultante foi utilizado de entrada para aplicação do algoritmo de agrupamento *k-means*[4]. O algoritmo utilizado está na classe de abordagem do problema de *K - Clussterização*, como explicado no Capítulo 2, dessa forma, fez-se necessário a escolha prévia da quantidade ideal do número de *clusters* que melhor se encaixam na base de dados.

A primeira métrica utilizada para avaliação da qualidade da clusterização foi o método do cotovelo, do inglês, *elbow method*. Segundo Bholowalia e Kumar (2014), o método se baseia na porcentagem de variância explicada em função do número de clusters. A partir de um $k = 2$, é feita iteração para um valor de k crescendo em uma unidade por vez até um limite determinado. O número ideal será definido no momento em que o ganho de informação sofrer uma queda, formando um ângulo mais acentuado no gráfico. O ganho de informação é medido pela soma dos quadrados intra-clusters, do inglês *within-clusters sum-of-squares* (WCSS). Podemos ver o resultdo obtido pela aplicação do método na Figura 3.2.

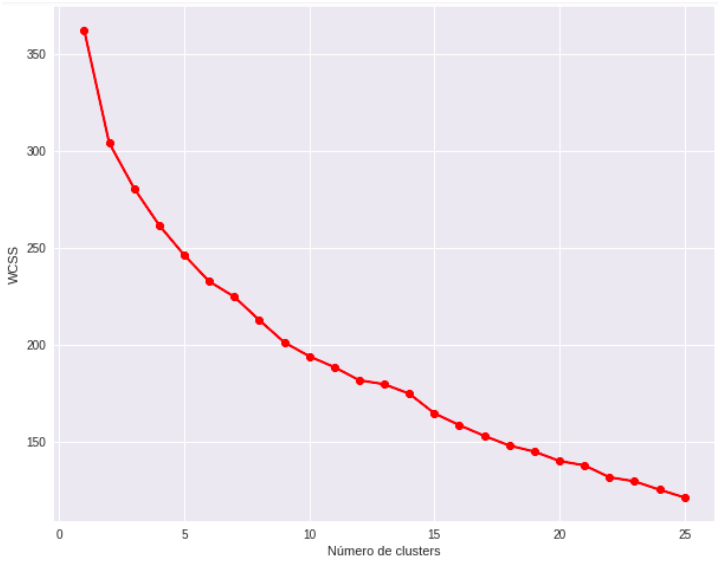


Figura 3.2: Método do Cotovelo

Posteriormente, utilizamos a métrica do *silhouette coefficient* para obtenção de uma medida mais efetiva e comparar com o método do cotovelo utilizado anteriormente. O coeficiente retornado pela métrica nos fornece informações da similaridade de uma observação em relação ao próprio *cluster* quando comparado aos outros *clusters* (Alvarez Orbe et al., 2021). A métrica pode ser descrita pela Equação 3.1.

[4] `https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html`

$$s(i) = \frac{b(i) - a(i)}{max(b(i), a(i))} \tag{3.1}$$

Onde $a(i)$ é a distância média entre o item i e todos os outros pontos em seu cluster e $b(i)$ é a distância média entre o item i e todos os outros pontos do cluster mais próximo (Silva et al., 2019). Analisando a Equação 3.1, observamos que para $a(i) >> b(i)$ teremos $s(i)$ se aproximando de -1, o que nos indica uma má classificação do item i. Por outro lado, para $b(i) >> a(i)$ teremos um coeficiente $s(i)$ se aproximando de 1, que nos mostra que o item i estará bem classificado.

Quantidade de *clusters*	**Média *Silhouette***
2	0.207280
3	0.105467
4	0.115659
5	0.119385
6	0.121821
7	0.112660
8	0.132207
9	0.138444
10	0.139793

Tabela 3.6: Valores da média *silhouette*

Dessa forma, foi retirado uma média de todos os coeficientes silhouette entre 2 a 10 *clusters*, como ilustrado na Tabela 3.6. De acordo as duas métricas utilizadas, foi escolhido 2 *clusters* como a quantidade ideal a ser utilizada para a pesquisa.

3.2.1 Análise dos resultados

Finalizada a etapa de mineração de dados com a rotulação de cada observação em seu respectivo *cluster*, iniciamos a interpretação e a avaliação dos resultados obtidos. Primeiramente, foi verificado a quantidade de observações em cada um dos grupos. Analisando a figura 3.3, podemos observar que 70.5% dos alunos foram agrupados no

cluster 0, correspondendo a exatas 93 observações, e 29.5% dos alunos sendo agrupados no *cluster* 1, equivalente a 39 observações.

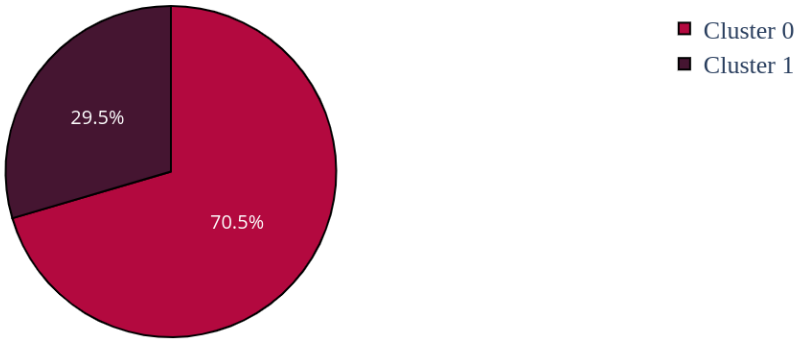


Figura 3.3: Quantidade de observações em cada *cluster*

Observando a Figura 3.4, notamos que o *cluster* 0, que contém maior parte das observações, possui também uma maior dispersão nas idades, com uma alta concentração em torno do primeiro quartil, 19 anos, e da mediana, 22 anos, além de que 75% dos alunos estando abaixo dos 25.25 anos. O *cluster* 1, no entanto, é caracterizado por uma menor dispersão, com uma maior concentração entre o primeiro e terceiro quartil, 20 e 23.75 anos, respectivamente.

Ao analisarmos o atributo período de evasão, Figura 3.5, podemos identificar um padrão distinto em relação as idades de evasão. O *cluster* 0 é caracterizado por uma alta concentração de evasão nos períodos iniciais dos cursos, entre o primeiro e segundo período, sendo representados pelo primeiro quartil e mediana no gŕafico, respectivamente. O *cluster* 1 no entanto é formado com uma maior dispersão nestes valores, além de uma maior frequência nos períodos mais tardios. Observamos uma clara concentração no intervalo entre o primeiro quartil, segundo período, até o terceiro quartil, quinto período.

Observamos também como os *clusters* foram divididos de forma que os alunos que apresentaram problemas relacionados ao curso ou universidade foram agrupados no *cluster* 1, ao passo que no *cluster* 0, pequenas porções das observações relataram tais problemas. A Figura 3.6 nos mostra que apenas 9.68% dos alunos do *cluster* 0 justificaram o baixo rendimento nas disciplinas como motivo de evasão, enquanto no *cluster* 1, este

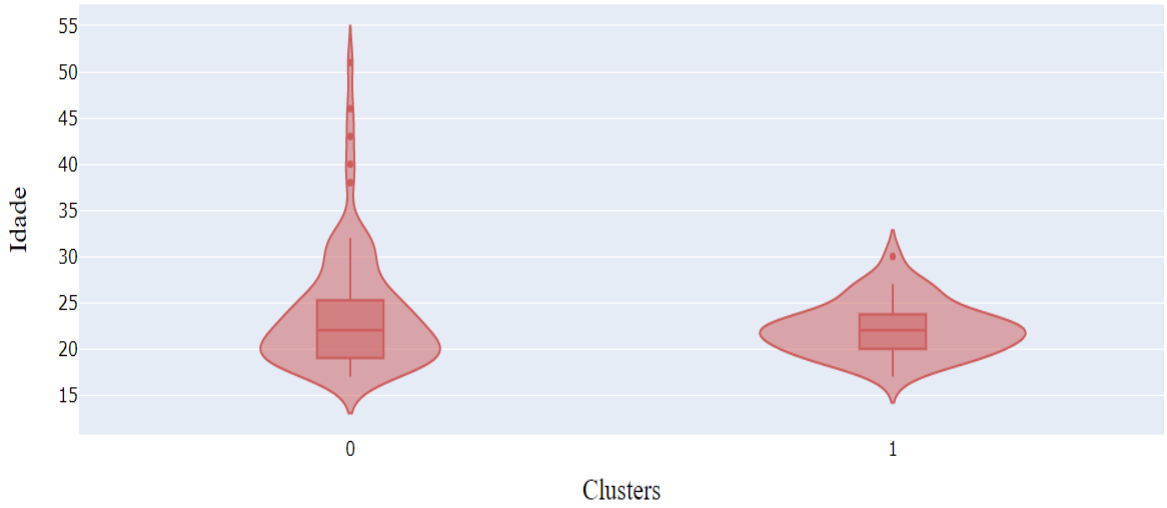


Figura 3.4: Distribuição das idades em cada *cluster*

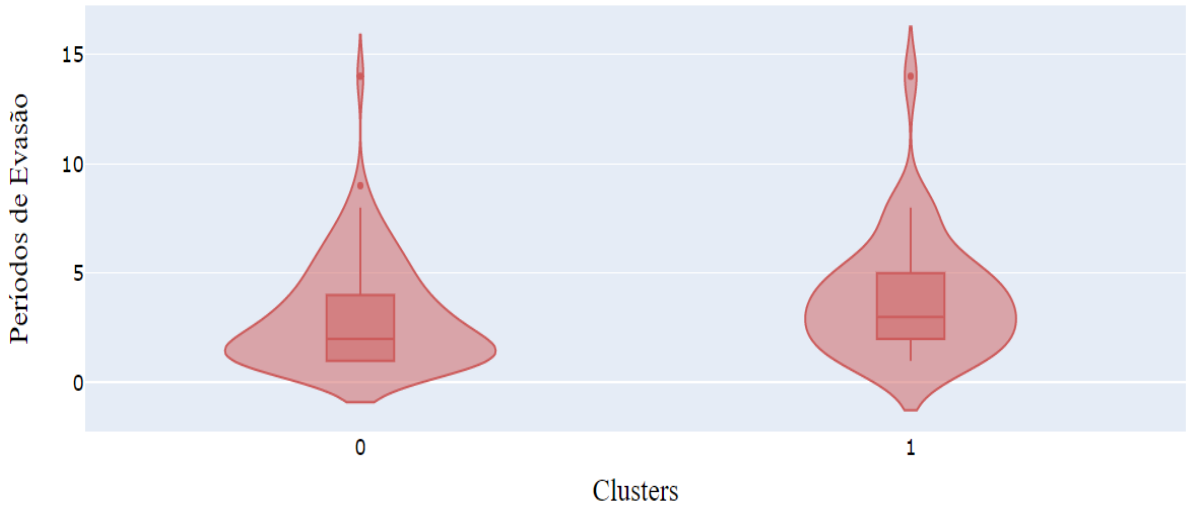


Figura 3.5: Distribuição dos períodos de evasão em cada *cluster*

problema é relatado por 84.6% dos alunos. Seguindo a mesma lógica, vemos pela Figura 3.7 que 69.2% dos alunos do *cluster* 1 tiveram problemas de adaptação com o curso como motivo de evasão, de forma que no *cluster* 1 essa taxa chegou a apenas 17.2%.

Dos alunos agrupados no *cluster* 1, de acordo a Figura 3.8, 66.7% apresentaram dificuldade na áera escolhida, e no *cluster* 0, o percentual chegou a apenas 6.45%. A partir da Figura 3.9, notemos que 59% dos alunos dos *cluster* 1 tiveram problemas de aprendizagem em relação ao conteúdo, e somente 4.3% das observações do *cluster* 0 relataram este problema.

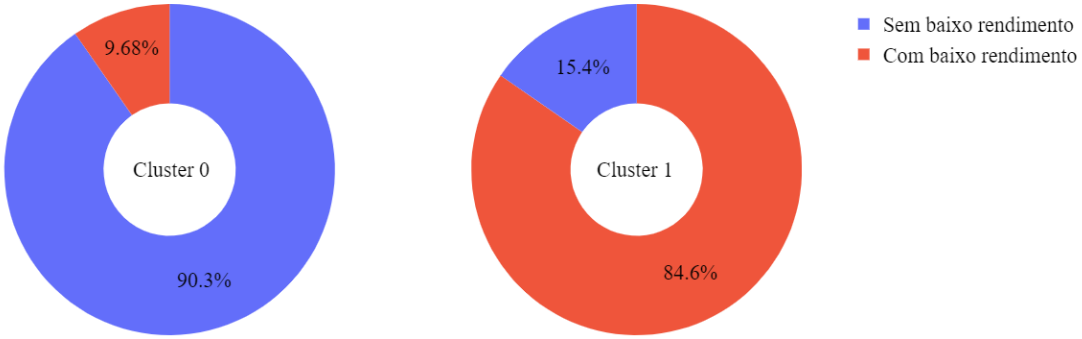


Figura 3.6: Baixo rendimento nas disciplinas em cada *cluster*

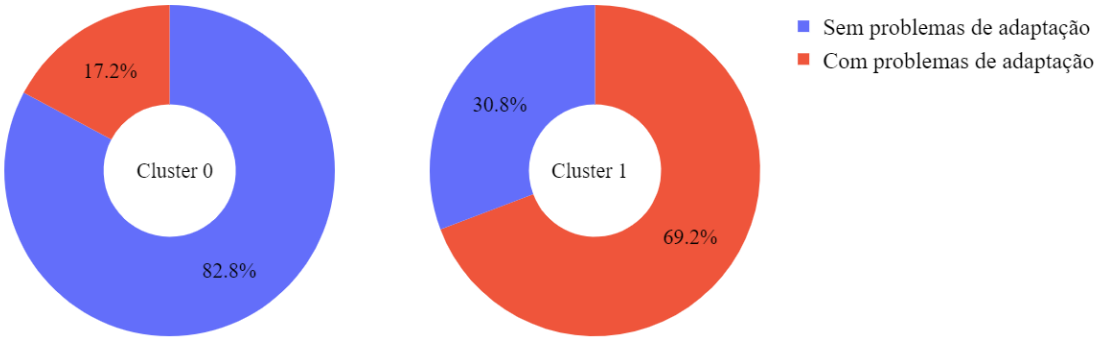


Figura 3.7: Problemas de adaptação com o curso em cada *cluster*

Como possível consequência da alta taxa de relatos de problemas desenvolvidos diretamente relacionados ao curso ou universidade, está a discrepância de problemas psicológicos em cada um dos grupos, ilustrado pela Figura 3.10, houve apenas uma pequena fração de ocorrências no *cluster* 0, 10.8%, enquanto o problema chegou a ser relatado por quase metade do *cluster* 1, 48.7%. A Figura 3.11 nos permite vizualizar como a distribuição dos sexos entre os dois grupos formados possui uma maior proximidade, sendo o *cluster* 0 composto por 74.2% do sexo masculino e 25.8% do sexo feminino, enquanto o *cluster* 1 formado por 33.3% do sexo feminino e 66.7% do sexo masculino.

Obsevando a Figura 3.13, notamos como as ocorrências de distanciamento da família

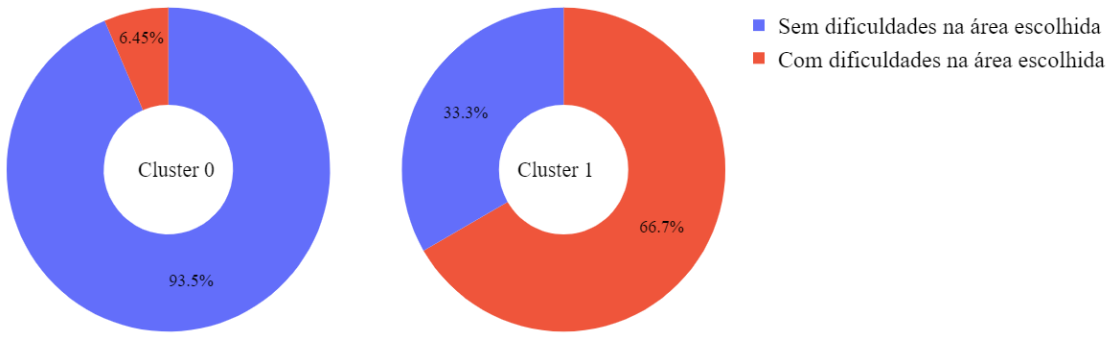


Figura 3.8: Dificuldade na área escolhida em cada *cluster*

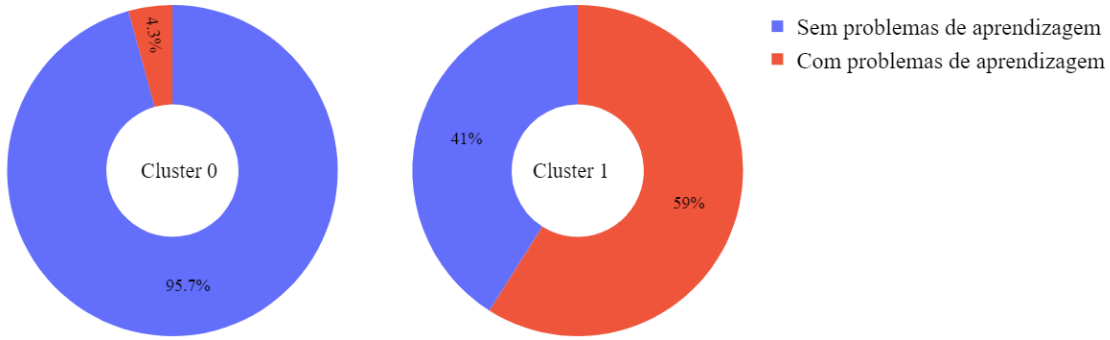


Figura 3.9: Problemas de aprendizagem em relação ao conteúdo em cada *cluster*

obteve uma divisão próxima entre os grupos, sendo alegada como motivo de evasão por 11.8% no *cluster* 0 e 17.9% no *cluster* 1. De maneira análoga aos problemas anteriomente analisados, os problemas financeiros teve uma maior taxa de relatos no *cluster* 1, correspondendo a 30.8%, enquanto no *cluster* 0 as alegações chegaram a apenas 7.53%.

Seguindo um padrão distinto, o atributo de oportunidade de ingresso em outra faculdade com mesmo curso possui maiores alegações no *cluster* 0, com um total de relatos de 24.7%, de forma que que no *cluster* 1, apenas 2.56% dos alunos tiveram tais oportunidades em outra faculdade.

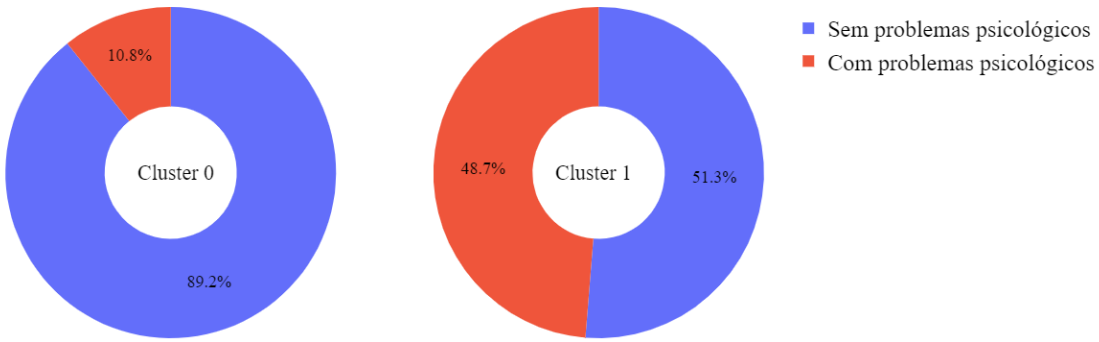


Figura 3.10: Problemas psicológicos em cada *cluster*

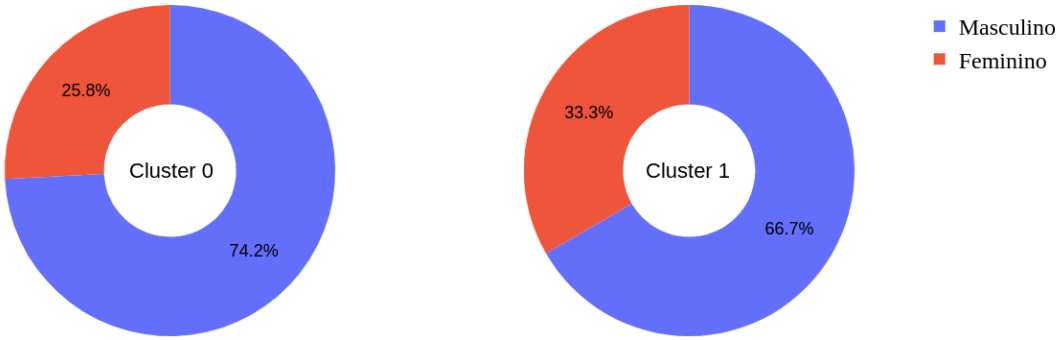


Figura 3.11: Distribuição dos sexos em cada *cluster*

Os sentimentos relacionados a cada aluno com o seu texto de motivo de evasão foi análisado de acordo a média em cada *cluster*, como podemos ver na Figura 3.12. O *cluster* 0 possui uma média de sentimentos positivo mais alta em relação ao *cluster* 1, tanto para a análise realizada no português quanto no inglês. O sentimento negativo é maior em magnitude no *cluster* 1, ou seja, os textos dos alunos desse grupo estavam mais carregados de sentimento negativo quando comparado ao *cluster* 0, o que pode ser justificável pelas análies anteriores, observando o padrão de maiores parcelas de alegação de diferentes tipos de problemas que ocasionaram a evasão no *cluster* 1.

Para fins de observações gerais, alguns atributos de interesse foram analisados mesmo

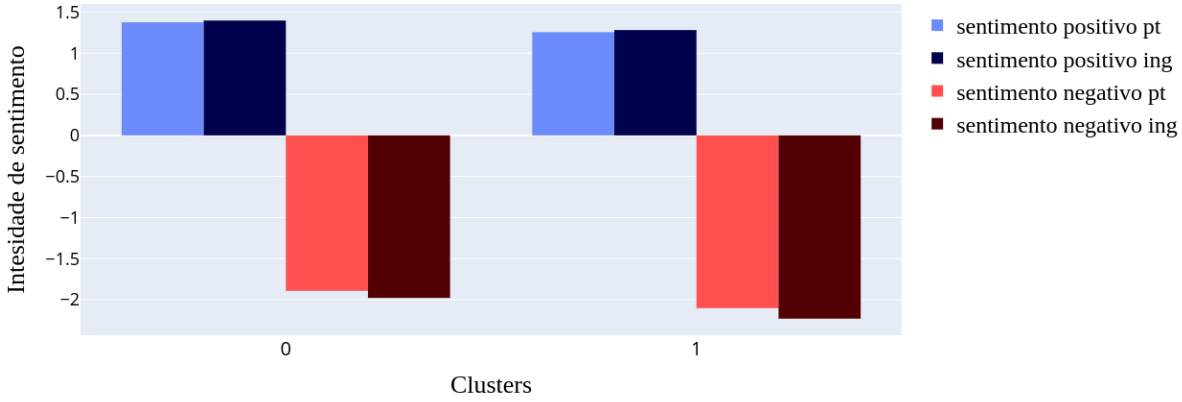


Figura 3.12: Média dos sentimentos em cada *cluster*

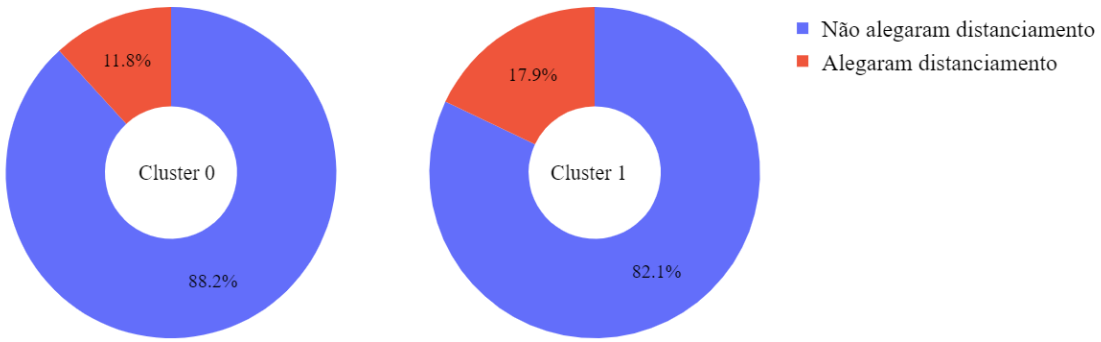


Figura 3.13: Distanciamento da família em cada *cluster*

não fazendo parte do grupo de atributos mais importantes para a análise dos componentes principais. Foi verificado assim que o padrão de relatos de problemas relacionados a faculdade permanece com uma maior taxa de ocorrências no *cluster* 1. Primeiramente analisamos os problemas com professores específicos, Figura 3.16, sendo um dos motivos de evasão em 35.9% das observações do *cluster* 1 e 9.68% do *cluster* 0. Posteriormente, observamos as ocorrências de desinteresse com o curso escolhido, Figura 3.17, o *cluster* 0 teve uma porcentagem de 24.7% de alunos que relataram como motivo de evasão, ao passo que no *cluster* 1, o desinteresse foi motivo de evasão para 48.7% dos alunos.

Avaliando as oportunidades não relacionadas diretamente com problemas com a uni-

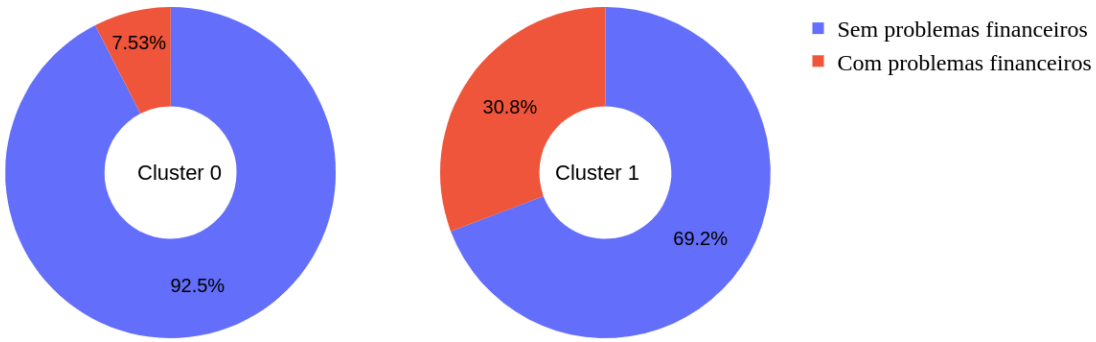


Figura 3.14: Problemas financeiros em cada *cluster*

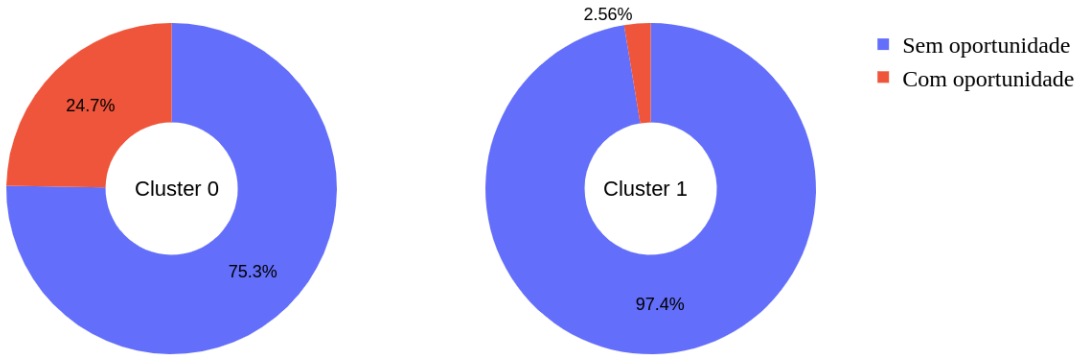


Figura 3.15: Oportunidades em outra faculdade com mesmo curso em cada *cluster*

versidade, a Figura 3.18 nos mostra como o *cluster* 0 possui uma maior porcentagem de relatos, 14%, quando comparado ao *cluster* 1, com 2.56% de alegações. Por último, examinamos o atributo de problemas com uma disciplina em específico, como ilustrado na Figura 3.19, uma pequena parcela de 8.6% do *cluster* 0 relatou o problema como motivo de evasão, no *cluster* 1 a parcela de ocorrência do problema chegou a 43.6% das observações totais.

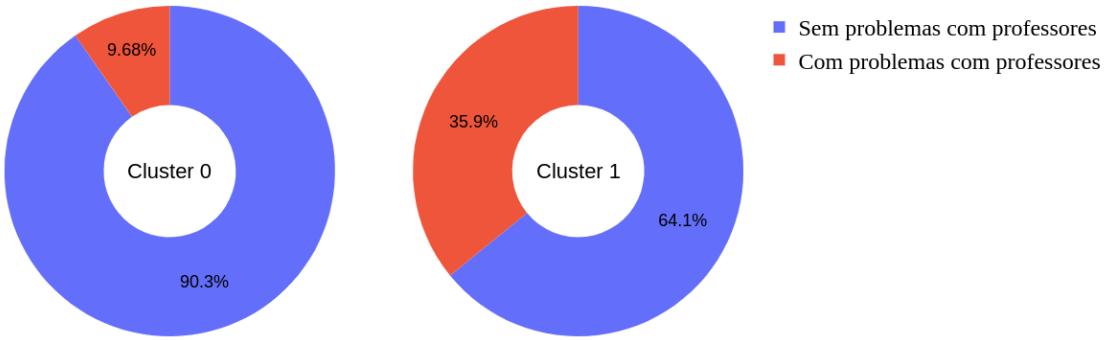


Figura 3.16: Problemas com professores específicos em cada *cluster*

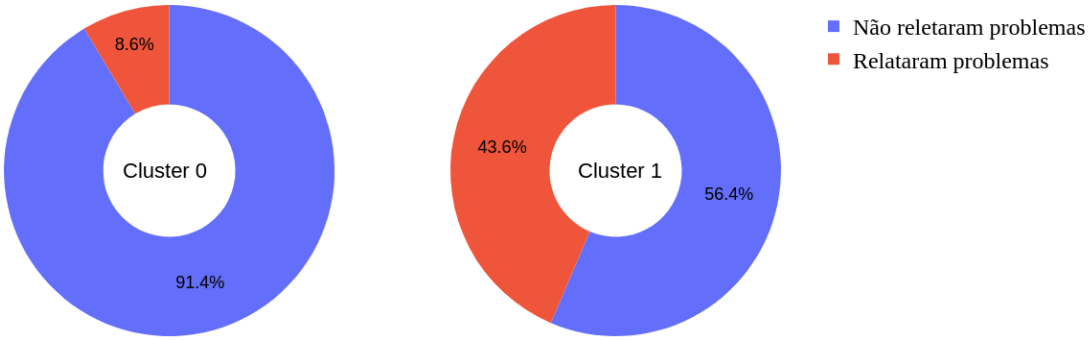


Figura 3.17: Desinteresse com curso escolhido em cada *cluster*

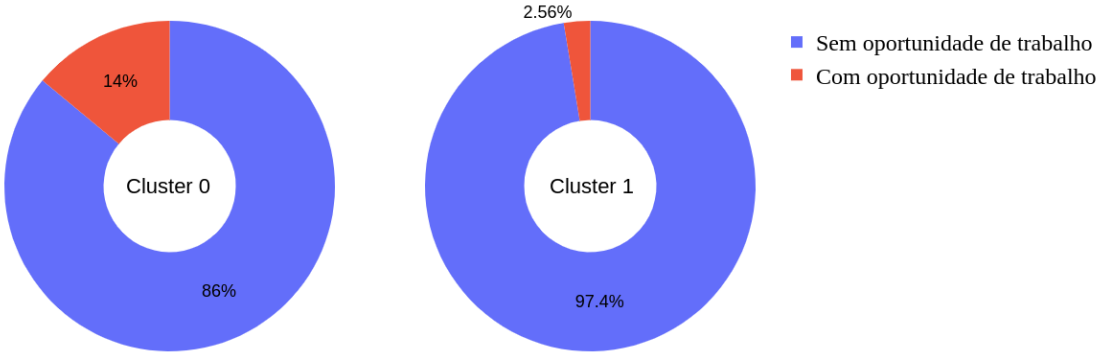


Figura 3.18: Oportunidade de trabalho em cada *cluster*

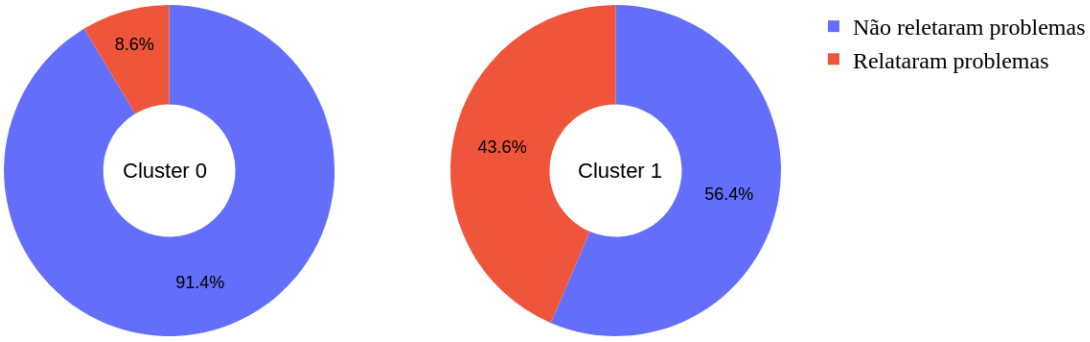


Figura 3.19: Problemas em uma disciplina espefíca em cada *cluster*

Capítulo 4

Conclusão e Trabalhos Futuros

Os padrões encontrados após a aplicação de todas as etapas do KDD, finalizando com os algoritmos para a mineração de dados e extração de conhecimento útil, evidenciaram a existência de duas classes de alunos que evadem a universidade, como vimos através da análise da clusterização.

O primeiro grupo de alunos, relativo ao *cluster* 0, possui uma maior concentração nas idades mais elevadas. Esse mesmo *cluster* de alunos é caracterizado por uma alta concentração de evasão em períodos iniciais dos cursos, Figura 3.5, tendo uma parcela de 50% das observações evadindo até o segundo período e 75% até o quarto período. Notamos também como pequenas porções relataram problemas ligados diretamente com o curso ou universidade, com exceção do desinteresse com o curso escolhido, sendo relatado por 24.7% dos alunos. Podemos associar tal observação através baixo sentimento negativo observado em suas frases, e por outro lado, um sentimento positivo maior, quando comparado ao *cluster* 1. Além das pequenas parcelas de problemas relatados como motivo de evasão, foi possível observar como os alunos deste grupo tiveram mais oportunidades externas ao campus de estudo, como oportunidades de trabalho e ingresso em outra faculdade com o mesmo curso, podendo ser associado os períodos de evasão no início do curso.

Já para o segundo grupo de alunos, relativos ao *cluster* 1, é caracterizado pela associação de alunos mais jovens, de forma que 75% estão abaixo dos 23.75 anos, Figura 3.4. O período de evasão possui maiores valores e mais distríbuídos quando comparado ao primeiro grupo, Figura 3.5. O aspecto marcante do *cluster* em questão foi a alta taxa de relatos de problemas ligados ao curso, disciplinas, adaptação ou com professo-

res específicos, podendo ser associado com a alta ocorrência dos problemas psicológicos, como vemos pela Figura 3.10, existindo a possibilidade de serem acarretados devido aos problemas abordados. Outra confirmação do mau relacionamento desevolvido com a universidade ser um motivo de evasão, está na média dos sentimentos analisados, Figura 3.12, os alunos agrupados neste *cluster* possui um sentimento positivo próximo da neutralidade, enquanto o sentimento negativo foi observado com uma maior força.

Através da análise das correlações positivas, identificamos como o baixos rendimentos podem levar o aluno ao desligamento, através das relação entre as variáveis de desligamento por reprovações em todas as disciplinas dois semestres consecutivos e dseligamento por coeficiente abaixo de 3 dois semestres consecutivos. Outra correlação observada está em como um aumento na dificuldade na área escolhida ou problemas com aprendizado em relação ao conteúdo podem acarretar em um baixo rendimento nas disciplinas. Por fim, aumentos nos problemas de adaptação com o curso ou problemas com aprendizado em relação ao conteúdo podem aumentar também a dificuldade na área escolhida.

As correlações negativas evidenciam como um aumento nas oportunidades de ingresso em outra faculdade com mesmo curso podem impactar em um decréscimo nos atributos de baixo rendimento nas disciplinas, dificuldade na área escolhida, problemas de adaptação com o curso e desinteresse com o curso escolhido, dessa forma, podemos avaliar de modo que se um aluno possui oportunidades em um mesmo curso em outra universidade, seu rendimento no mesmo curso na universidade atual deve ser satisfatório, ou do contrário, o aluno não buscará por tais oportunidades. Outras correlações a serem destacadas são a forma com que aumentos em problemas com professores específicos ou nos períodos de evasão, podem implicar em maiores sentimentos negativos no aluno.

A partir do trabalho desevoldido, diferentes novas pesquisas podem ser realizadas a fim de aprimoramento ou complemntação dos resultados alcançados, as quais podemos citar:

- Realização da pesquisa em um conjunto de dados maior.
- Aprimoramento e obtenção de melhores resultados para o agrupamento dos alunos.
- Cruzamento dos resultados obtidos com a vida acadêmica diária do aluno.
- Realizar a análise de sentimentos em *softwares* distintos para verificação dos resultados.

Referências Bibliográficas

Abdi, H. e Williams, L.: 2010, Principal component analysis, *Wiley Interdisciplinary Reviews: Computational Statistics* **2**, 433–459.

Alvarez Orbe, C. E. et al.: 2021, Benchmarking de desempenho energético de edifícios educacionais: uma abordagem de mineração de dados.

Alves Pinto, A. M.: 2019, Caracterização da ansiedade e análise de sentimentos dos alunos de computação do Instituto de Ciências Exatas e Aplicadas (ICEA)/Universidade Federal de Ouro Preto (UFOP). Monografia (Bacharel em Sistemas de Informação), Instituto de Ciências Exatas e Aplicadas (ICEA) - Universidade Federal de Ouro Preto (UFOP), João Monlevade, Brasil.

Balahur, A. e Turchi, M.: 2012, Multilingual sentiment analysis using machine translation?, *Proceedings of the 3rd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, pp. 52–60.

Benevenuto, F., Ribeiro, F. e Araújo, M.: 2015, Métodos para análise de sentimentos em mídias sociais, *Sociedade Brasileira de Computação* .

Bholowalia, P. e Kumar, A.: 2014, Ebk-means: A clustering technique based on elbow method and k-means in wsn, *International Journal of Computer Applications* **105**(9).

Borges, L. E.: 2014, *Python para desenvolvedores*, 1 edn, Novatech.

de Carvalho Melo Lobo, M. B.: 2012, Panorama da evasão no ensino superior brasileiro: aspectos gerais das causas e soluções, *ABMES* p. 1.

de Oliveira, R. F.: 2021, Análise de sentimentos das postagens e comentários dos principais candidatos à eleição presidencial brasileira de 2018. Monografia (Bacharel em Engenharia de Computação), Instituto de Ciências Exatas e Aplicadas (ICEA) - Universidade Federal de Ouro Preto (UFOP), João Monlevade, Brasil.

DIAS, E. C. M., THEÓPHILO, C. R. e LOPES, M. A.: 2010, Evasão no ensino superior: estudo dos fatores causadores da evasão no curso de ciências contábeis da universidade estadual de montes claros–unimontes–mg, *Congresso USP de Iniciação Científica em Contabilidade*, Vol. 7, pp. 1–16.

Doni, M. V.: 2004, Análise de cluster: Métodos hierárquicos e de particionamento.

Fayyad, U., Piatetsky-Shapiro, G. e Smyth, P.: 1996, From data mining to knowledge discovery in databases, *AI Mag.* **17**, 37–54.

Jain, A. K., Murty, M. N. e Flynn, P.: 1999, Data clustering: a review, *ACM Comput. Surv.* **31**, 264–323.

Jolliffe, I. T. e Cadima, J.: 2016, Principal component analysis: a review and recent developments, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **374**.

Kanyongo, G. Y.: 2005, Determining the correct number of components to extract from a principal components analysis: A monte carlo study of the accuracy of the scree plot, *Journal of Modern Applied Statistical Methods* **4**(1), 13.

Liu, B. e Zhang, L.: 2012, A survey of opinion mining and sentiment analysis, *Mining Text Data*.

Moreira, M. P. e Favero, E. L.: 2009, Um ambiente para ensino de programação com feedback automático de exercícios, *Workshop sobre Educação em Computação (WEI 2009)*, Vol. 17.

Mukaka, M.: 2012, Statistics corner: A guide to appropriate use of correlation coefficient in medical research., *Malawi medical journal : the journal of Medical Association of Malawi* **24 3**, 69–71.

Ochi, L. S., Dias, C. R. e Soares, S. S. F.: 2004, Clusterização em mineração de dados, *Instituto de Computação-Universidade Federal Fluminense-Niterói* **1**, 46.

Reis, J. C. S., Gonçalves, P., Araújo, M., Pereira, A. C. M. e Benevenuto, F.: 2016, Uma abordagem multilıngue para analise de sentimentos.

Salton, G.: 1991, Developments in automatic text retrieval, *Science* **253**, 974 – 980.

Schober, P., Boer, C. e Schwarte, L.: 2018, Correlation coefficients: Appropriate use and interpretation, *Anesthesia & Analgesia* **126**, 1763–1768.

Schwartz, H. A., Eichstaedt, J., Kern, M. L., Dziurzynski, L., Ramones, S., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. e Ungar, L.: 2013, Personality, gender, and age in the language of social media: The open-vocabulary approach, *PLoS ONE* **8**.

Silva, D. E., Bittencourt, R. A. e Calumby, R. T.: 2019, Clustering similarity measures for architecture recovery of evolving software, *Anais do VII Workshop on Software Visualization, Evolution and Maintenance (VEM)*, SBC, pp. 45–52.

Thelwall, M.: 2013, Heart and soul: Sentiment strength detection in the social web with sentistrength.

Xu, F., Kurz, D., Piskorski, J. e Schmeier, S.: 2002, A domain adaptive approach to automatic acquisition of domain relevant terms and their relations with bootstrapping, *LREC*.

Zacarias, T. C.: 2019, Sistema de modelagem de perfis de usuários baseados em mídias sociais. Monografia (Bacharel em Engenharia de Computação), Instituto de Ciências Exatas e Aplicadas (ICEA) - Universidade Federal de Ouro Preto (UFOP), João Monlevade, Brasil.