
**Universidade Federal de Ouro Preto**
**Instituto de Ciências Exatas e Aplicadas**
**Departamento de Computação e Sistemas**

# Caracterização do problema de evasão de discentes nos cursos do ICEA mediante técnicas de mineração de dados

**Danilo Martins Caldeira**

## TRABALHO DE CONCLUSÃO DE CURSO

ORIENTAÇÃO:
Janniele Aparecida Soares Araújo

COORIENTAÇÃO:
Helen de Cássia Sousa da Costa Lima

**Setembro, 2021**
**João Monlevade–MG**

**Danilo Martins Caldeira**

# Caracterização do problema de evasão de discentes nos cursos do ICEA mediante técnicas de mineração de dados

Orientador: Janniele Aparecida Soares Araújo

Coorientador: Helen de Cássia Sousa da Costa Lima

Monografia apresentada ao curso de Sistemas de Informação do Instituto de Ciências Exatas e Aplicadas, da Universidade Federal de Ouro Preto, como requisito parcial para aprovação na Disciplina "Trabalho de Conclusão de Curso II".

**Universidade Federal de Ouro Preto**

**João Monlevade**

**Setembro de 2021**

SISBIN - SISTEMA DE BIBLIOTECAS E INFORMAÇÃO

C146c Caldeira, Danilo Martins.
Caracterização do problema de evasão de discentes nos cursos do ICEA mediante técnicas de mineração de dados. [manuscrito] / Danilo Martins Caldeira. - 2021.
65 f.: il.: color., gráf., tab..

Orientadora: Profa. Dra. Janniele Aparecida Soares Araújo.
Coorientadora: Profa. Ma. Helen de Cassia Sousa da Costa Lima.
Monografia (Bacharelado). Universidade Federal de Ouro Preto. Instituto de Ciências Exatas e Aplicadas. Graduação em Sistemas de Informação .

1. Mineração de dados. 2. Descoberta de conhecimento em base de dados. 3. Evasão universitária. I. Araújo, Janniele Aparecida Soares. II. Lima, Helen de Cassia Sousa da Costa. III. Universidade Federal de Ouro Preto. IV. Título.

CDU 004.62:378

Bibliotecário(a) Responsável: Sione Galvão Rodrigues - CRB6 / 2526

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIENCIAS EXATAS E APLICADAS
DEPARTAMENTO DE COMPUTACAO E SISTEMAS

**FOLHA DE APROVAÇÃO**

**Danilo Martins Caldeira**

**Caracterização do problema de evasão de discentes nos cursos do ICEA mediante técnicas de mineração de dados**

Monografia apresentada ao Curso de Sistemas de Informação da Universidade Federal
de Ouro Preto como requisito parcial para obtenção do título de Bacharel em Sistemas de Informação

Aprovada em 02 de setembro de 2021

Membros da banca

Doutora - Janniele Aparecida Soares Araujo - Orientador(a) - Universidade Federal de Ouro Preto
Mestra - Helen de Cassia Sousa da Costa Lima - Coorientadora - Universidade Federal de Ouro Preto
Doutor - Fernando Bernardes de Oliveira - Universidade Federal de Ouro Preto
Doutora - Lucineia Souza Maia - Universidade Federal de Ouro Preto

Janniele Aparecida Soares Araujo, orientadora do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 13/09/2021

Documento assinado eletronicamente por **Janniele Aparecida Soares Araujo**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 13/09/2021, às 15:53, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **0220127** e o código CRC **C92AAD70**.



R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35400-000
Telefone: - www.ufop.br

*Este trabalho é dedicado à minha mãe, Solange Martins e ao meu pai Zito Caldeira. Eles foram meu maior apoio. Posso dizer com firmeza que não é apenas uma realização pessoal a finalização da minha graduação, mas é um mérito coletivo, meu e dos meus pais.*

# Agradecimentos

Podemos fazer uma analogia de algumas fases da nossa vida ao conceito de Sistema de Informação, um conjunto de componentes inter-relacionados formando o todo. Para alcançar a conclusão deste trabalho, contei com apoio de várias pessoas. Quero agradecer a todos que me apoiaram e ajudaram a chegar neste ponto.

Agradeço primeiramente a Deus, pela minha vida e as oportunidades a mim concedidas. Pela força que sempre encontrei Nele para vencer meus desafios, poder dedicar o melhor de mim e nunca desistir quando houveram obstáculos, e sim, aprender a solucioná-los.

Aos meus pais, Zito Caldeira e Solange Martins, que são os meus principais pilares para tudo que sou hoje e para tudo aquilo que conquistei e tenho conquistado. Eles, muitas vezes, se desdobraram para ajudar-me nas realizações dos meus objetivos. Minha mãe, apenas com ensino fundamental, tem um conhecimento impressionante e é minha maior fonte de inspiração. Ela sempre teve o sonho de estudar, por falta de oportunidade não continuou, mas sempre motivou-me e se esforçou para que eu pudesse. Meu pai, marceneiro desde adolescente, nunca teve o objetivo estudar, mas sempre via o meu interesse, apoiando, estimulando e fazendo tornar possível. Ao meu irmão, que sempre esteve do meu lado e mesmo sendo irmão mais novo, foi muitas vezes inspiração para mim. Agradeço a toda minha família, as minhas primas e primos, tios e tias que sempre estiveram ao meu lado. Agradeço imensamente a minha avó Aparecida Magalhães pelo seu carinho até hoje e a minha avó Maria das Dores (*in memorian*), que mesmo não estando mais ao meu lado, foi uma avó presente, que amava contar histórias, aprender sobre as novas coisas que eu estava estudando e agregando o meu saber com suas experiências nas nossas conversas.

Agradeço a todos os amigos, que estiveram ao lado nos dias de comemorações, dias normais e nos dias difíceis. Amigos de infância, amigos da escola, alguns que hoje não tenho muito contato, mas que representaram muito na minha vida. Aos amigos que a UFOP trouxe, pessoas incríveis que sempre pude contar, que sempre estavam juntos comigo, todos amigos do Laboratório de Engenharia e Desenvolvimento de Sistemas (LEDS) e aos novos amigos que estou tendo oportunidade de conhecer por ter mudado de cidade e começado minha carreira profissional.

Agradeço imensamente a todos os professores que tive na universidade, que se dedicam, transformando alunos em profissionais capacitados, pessoas capazes e qualificadas. Agradeço ao professor Fernando Borges que me ensinou a programação inicial, ao que me interessei muito e sua disciplina influenciou muito a continuar no curso. A professora Daniela Dias, uma pessoa incrível que sempre apresentou muito bem como é o trabalho

dos profissionais formados em Sistemas de Informação. Ao professor e amigo Tiago Lima, que esteve ao meu lado, por quatro anos, orientando-me nos projetos de extensão no LEDS. Ao professor da área de Engenharia de Software Euler Horta, que agregou muito meu conhecimento na área. A Lucinéia Maia que me ajudou muito a decidir qual tema me dedicar no Trabalho de Conclusão. E especialmente, a professora Janniele Soares, por sua orientação, atenção e oportunidade de desenvolvermos juntos este trabalho. Em sua disciplina eu conheci a área de análise de dados e sistemas voltados ao apoio às tomadas de decisões, ao qual me despertou muito interesse. Através dela, tive a oportunidade de realizar a pesquisa neste ramo e trabalhar atualmente nesta área.

*"O homem não teria alcançado o possível se, repetidas vezes, não tivesse tentado o impossível."*

— Max Weber (1864 – 1920)

# Resumo

Este trabalho utiliza técnicas de mineração de dados em busca da caracterização do problema de evasão de alunos nos cursos do Instituto de Ciências Exatas e Aplicadas (ICEA). Com a utilização destas técnicas espera-se a obtenção de novos conhecimentos úteis relacionados à temática da evasão. A metodologia aplicada neste trabalho é baseada no processo *Knowledge Discovery in Databases* (KDD) e as etapas consistem em selecionar as fontes de dados, preprocessar, transformar e enriquecer, aplicar os algoritmos e analisar os resultados obtidos. Entre os resultados obtidos, destacam-se a criação de grupos reunindo alunos com similares e a percepção de alunos matriculados com características parecidas com alunos que vieram a evadir. Também foi possível a detecção de alunos anômalos em relação ao padrão.

**Palavras-chaves**: evasão. mineração de dados. processo de descoberta de conhecimento.

# Abstract

This work uses data mining techniques in order to understand the problem of the dropout rate of students in ICEA courses. With the use of these techniques, it is expected to obtain new useful knowledge related to the theme of dropout. The methodology applied in this work is based on the KDD process and the steps consist of selecting the data sources, preprocessing, transforming and enriching, applying the algorithms and analyzing the obtained results. Among the results obtained, we highlight the creation of groups bringing together students with similar types and the perception of students enrolled with similar characteristics to students who dropped out. It was also possible to detect students who were out of line with the pattern.

**Key-words**: evasion. data mining. knowledge discovery in databases.

# Lista de ilustrações

# Lista de tabelas

# Lista de abreviaturas e siglas

**API** *Application Programming Interface*

**CPF** Comprovante de Situação Cadastral

**EDM** *Educational Data Mining*

**ENADE** Exame Nacional de Desempenho dos Estudantes

**IBGE** Instituto Brasileiro de Geografia e Estatística

**ICEA** Instituto de Ciências Exatas e Aplicadas

**ICEB** Instituto de Ciências Exatas e Biológicas

**IES** Instituições de Ensino Superior

**IFMA** Instituto Federal do Maranhão

**KDD** *Knowledge Discovery in Databases*

**LGPD** Lei Geral de Proteção de Dados Pessoais

**NACE** Núcleo de Assuntos Comunitários Estudantis

**PCA** *Principal Component Analysis*

**SESU/MEC** Secretaria de Educação Superior do Ministério da Educação

**UFOP** Universidade Federal de Ouro Preto

**UFV** Universidade Federal de Viçosa

# Sumário

# 1 Introdução

A evasão é apontada por Veloso and de Almeida (2013) como um processo complexo e comum, presente nas Instituições de Ensino Superior (IES) e nos últimos anos esse tema tem sido foco em várias pesquisas.

Trabalhos relacionados à temática da evasão foram intensificados a partir de 1995, segundo Bardagi (2007), devido a criação da Comissão Especial de Estudos sobre Evasão, através da Secretaria de Educação Superior do Ministério da Educação (SESU/MEC), indicando uma valorização e maior preocupação política em relação ao problema. Porém, Bardagi (2007) comenta que o tema ainda não foi tratado com rigor e empenho necessários para seu entendimento.

O campus João Monlevade, inaugurado em 22 de setembro de 2002, é um dos institutos da Universidade Federal de Ouro Preto (UFOP), localizado no campus de João Monlevade. Iniciou-se oferecendo apenas o curso de Engenharia de Produção. No primeiro semestre do ano de 2005 passou a oferecer o curso de Sistemas e Informação. Já no segundo semestre de 2009, as graduações em Engenharia de Computação e Engenharia Elétrica foram incluídas.

Segundo dados obtidos através da Seção de Ensino do ICEA, departamento responsável pelo controle das matrículas, dados e históricos do instituto, o total de discentes que abandonaram a graduação no Instituto, desde a sua fundação a março de 2021 equivale a 2.386. Se comparado com o total de matrículas homologadas neste mesmo período, equivalente a 4.767, tem-se uma porcentagem de 50,05% de evasão em relação ao total de ingressantes. Para a direção, compreender, mensurar, prever e analisar a evasão são necessidades atuais. Percebe-se uma grande necessidade de conhecer o problema, ter um direcionamento maior para aplicações de medidas preventivas e corretivas e um melhor apoio à tomada de decisão. Assim, todos os resultados desta pesquisa serão repassados à diretoria do instituto, tendo-os como suporte para trabalhos e ações direcionadas ao problema.

O processo de descoberta de conhecimento em bases dados, incluindo a etapa de mineração de dados, é defendida por (Castanheira, 2008) como uma ferramenta usada amplamente para auxilio à tomada de decisão. Atualmente, trabalhos como dos autores Castanheira (2008); DIAS et al. (2010); Machado et al. (2015); Belenke dos Santos (2021) e vários outros vêm surgindo, onde aplicam-se as técnicas de mineração de dados para obtenção de conhecimento útil. Atualmente, segundo apresentado por Belenke dos Santos (2021), qualquer trabalho que utiliza-se de técnicas de mineração de dados voltados à área educacional, é classificado como *Educational Data Mining* (EDM).

Tendo em vista a gravidade do problema da evasão presente nas IES, a necessidade de entendimento acerca da problemática da evasão no ensino superior, o alto índice de evasão no ICEA, segundo os dados fornecidos pela seção de ensino, além dos ganhos e resultados promissores da utilização das técnicas de mineração de dados foram definidos os objetivos desta pesquisa, apresentado na seção seguinte.

## 1.1 O problema

Nesta seção são apresentados gráficos e indicadores obtidos em uma análise quantitativa dos dados disponibilizados pela Seção de Ensino, setor que é responsável pelo atendimento aos discentes dos cursos de graduação. Está descrito, de maneira detalhada, o problema que o ICEA vem enfrentando com a evasão. Esta análise inicial foi fundamental ao entendimento do problema abordado e percepção da gravidade, reforçar à justifica da pesquisa e proporcionar norteamento das tratativas a serem trabalhadas no desenvolvimento deste trabalho.

Evasão 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 ocorridos nas últimas três décadas, segundo Junior et al. (2019).

O índice de evasão considerado neste trabalho consiste na relação entre o total de discentes que evadiram no instituto e o total de discentes que ingressaram na instituição. O número total de discentes ingressados nos quatro cursos oferecidos pelo instituto, desde a sua fundação até maio de 2021, é de 4767 discentes e deste total, há 2386 discentes que se desligaram da instituição sem a conclusão da graduação. Isto significa que o índice geral de evasão do instituto é equivalente a 50,05%.

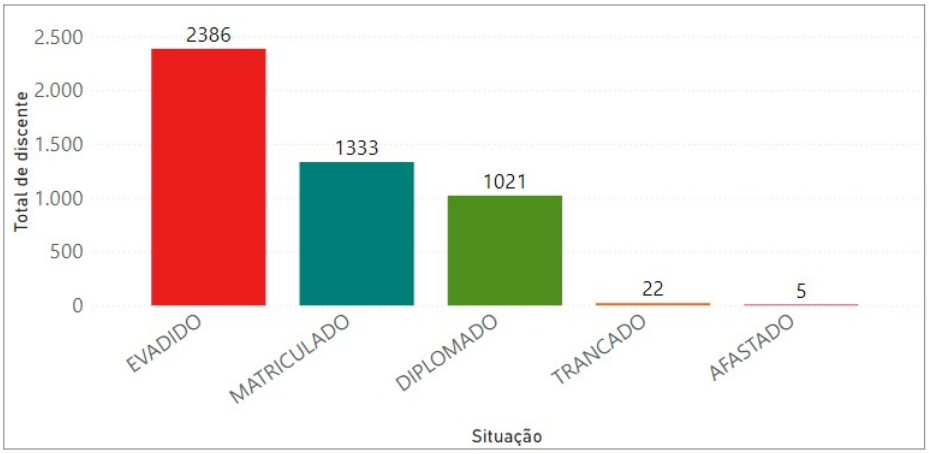
É fundamental a comparação do índice de evasão com outros dois importantes indicadores. O primeiro é o índice de diplomação, onde considera-se a relação entre o total de discentes formados em um ou mais cursos e o total de ingressantes. Entre 2002 e maio de 2021, houveram 1021 discentes diplomados no instituto, sendo o índice de diplomação equivalente a 21,41%. Percebemos um contraste significante entre o índice de evasão e o índice de diplomação. O segundo indicador é a taxa de discentes vinculados à instituição. Esta taxa equivale ao somatório dos discentes matriculados, em mobilidade acadêmica, discentes com matrículas trancadas e os afastados. Atualmente, esta taxa equivale a 1360 discentes. Este total precisa de uma atenção especial, pois representa os discentes que, no futuro, farão parte do grupo de diplomados ou evadidos.

Ao comparar os três indicadores, é possível observar que, aproximadamente a cada dez discentes que ingressam no ICEA, cinco evadiram, dois concluíram sua graduação e três encontram-se ainda vinculados à um curso na instituição. Buscando compreender melhor o histórico dos dados, será exposto a seguir gráficos e dados estatísticos que descreverão o

ingresso e a evasão dos discentes no instituto.

A Figura 1 apresenta a quantidade de discentes por cada tipo de situação.

Figura 1 – Total de discentes por situação

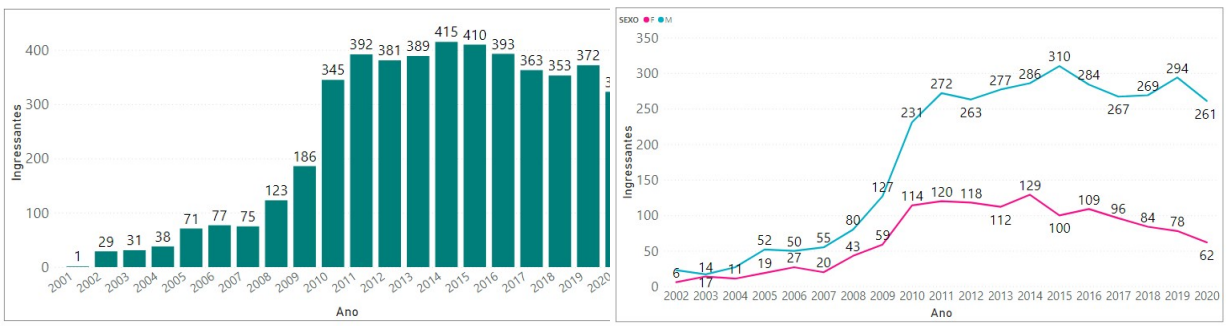



Fonte: Autor do trabalho

Iniciando com as análises referentes aos ingressos, a Figura 2a apresenta o total de ingressos entre os anos de 2002 a 2020. Deste total de ingresso, 27,73% são do sexo feminino e 72,27% são do sexo masculino. Observa-se que a maioria dos discentes que ingressam na instituição são do sexo masculino e um acompanhamento da distribuição anual por sexo é exposto na Figura 2b. A distribuição de ingressantes por curso é de 29,91% para o curso de Engenharia de Produção, com a maior representatividade, 25,49%, 22,84%, 22,36% respectivamente para os cursos de Sistemas de Informação, Engenharia Elétrica e Engenharia de Computação.

Figura 2 – Total de ingressos

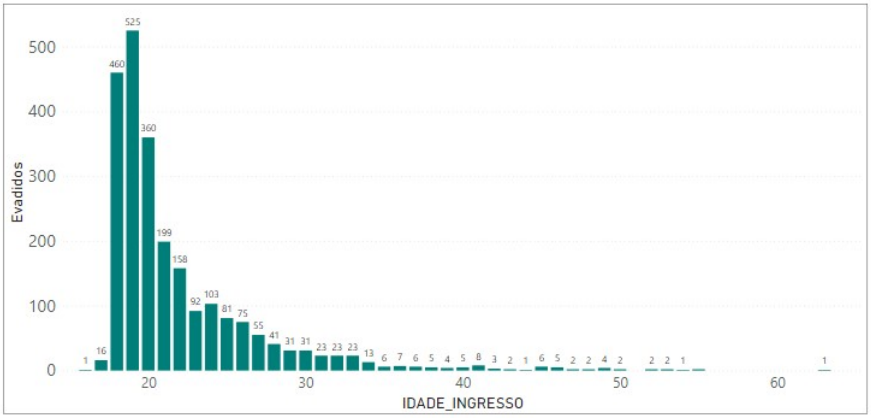
(a) Total de ingressos por ano (b) Total de ingressos por sexo e ano




Fonte: Autor do trabalho

Outra análise fundamental é acompanhar a faixa etária dos ingressantes. A Figura 3 apresenta este acompanhamento, mostrando a quantidade de discentes por faixa etária. Percebe-se que maioria dos discentes ingressaram com 18 a 22 anos e que, é menor o ingresso de discentes com idade superior a 35 anos.

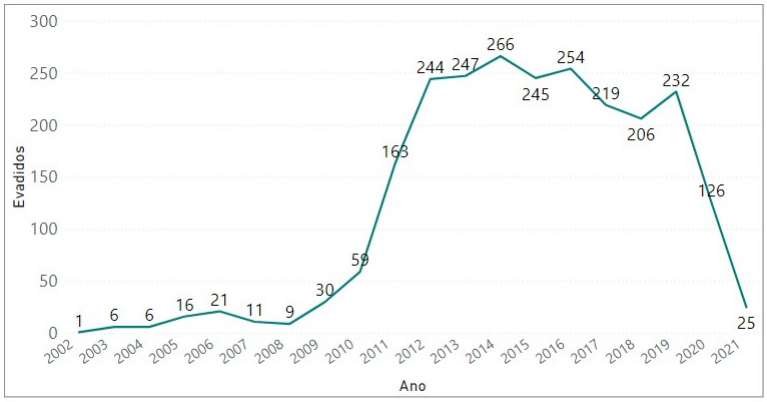
Figura 3 – Total de ingressos por idade




Fonte: Autor do trabalho

As análises a seguir são referentes às evasões ocorridas no instituto. A Figura 4 mostra a quantidade de evasões ocorridas por ano. Destaca-se o ano de 2014 como o ano que possui o maior número de evasões, com 266 ao total e uma queda significativa de evasões em 2020, em relação aos anos anteriores. Ressaltando que em 2020 a UFOP foi paralisada devido a pandemia do *Covid* 19.

Figura 4 – Total de evasões por ano



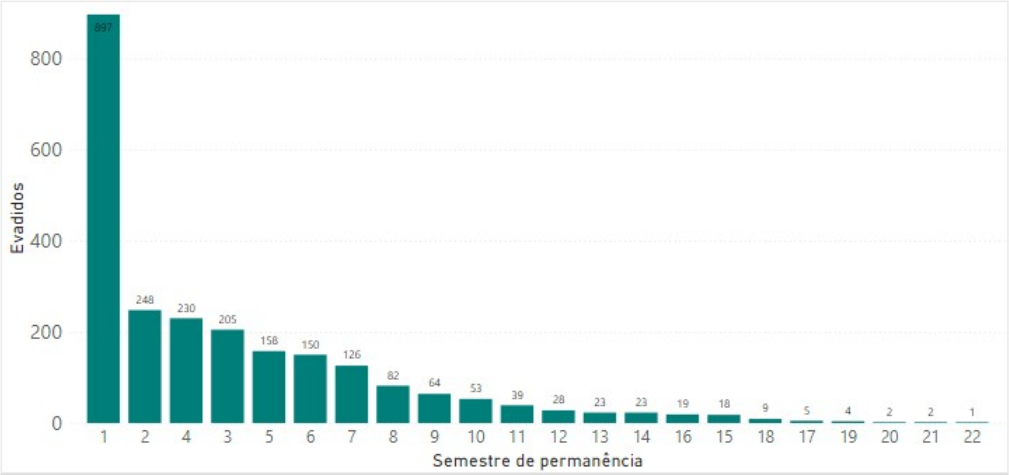

Fonte: Autor do trabalho

A evasão é maior no curso de Engenharia da Computação, com 28% do total de evasões e o menor índice está no curso de Engenharia de Produção, com 20% do total.

75% dos discentes evadidos são do sexo masculino e 25% do sexo feminino. 41 discentes, equivalente a 2% do total de evadidos, antes de evadirem da instituição, realizaram transferência interna, mudando de curso dentro do instituto. Analisando o percentual de discentes evadidos por sexo em relação ao total de ingressos por sexo, tem-se 46% para sexo femininos, onde, dos 1322, existem 611 discentes evadidos do sexo feminino. Já para o sexo masculino, temos 52% referentes aos 1775 discentes evadidos do total de 3445 ingressos.

Acompanhar a faixa de tempo que os discentes evadidos permaneceram na instituição é interessante. Para isto foi criado um fator estimando quantos semestres cada aluno ficou vinculado à UFOP. Neste cálculo, é realizado a subtração do ano de evasão pelo ano de ingresso, levando em considerações os casos onde os discentes ingressam ou evadem no meio do ano, obtendo o total de semestres de permanência. Este acompanhamento da permanência é apresentado na Figura 5. Com apenas um semestre de permanência, estão os discentes que abandonaram o curso no primeiro semestre e os discentes que homologaram a matrícula, mas não deram continuidade no curso.

Figura 5 – Total de evasões por semestre de permanência dos discentes

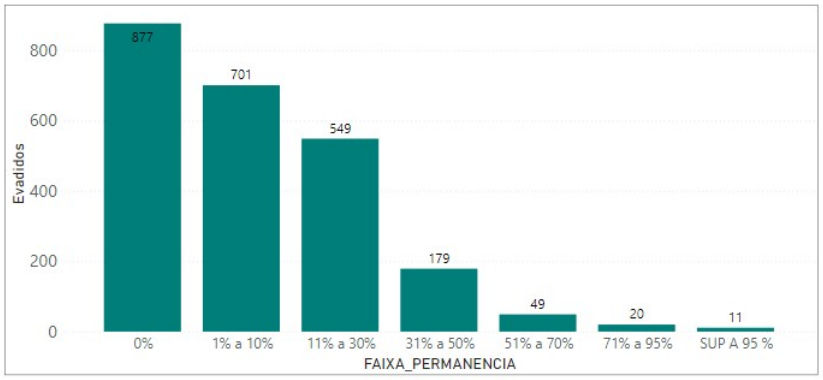



Fonte: Autor do trabalho

Além dos semestres de permanência, é interessante analisar as evasões em relação o percentual cursado. O cálculo deste percentual condiz entre a razão do total de horas cursadas por cada aluno e a carga horária do seu curso. Para facilitar análise, faixas foram criadas, sendo elas: 0%; 1% a 10%; 11% a 30%; 31% a 50%; 51% a 70%; 71% a 95%; e, superior a 95%.

A Figura 6 refere-se ao total evasão ocorrida em relação a cada uma dessas faixas. Percebe-se que a maioria dos discentes evadem quando possuem uma carga horária inferior a 30%, ou seja, quanto maior o percentual cursado, menor é o total de evasão.

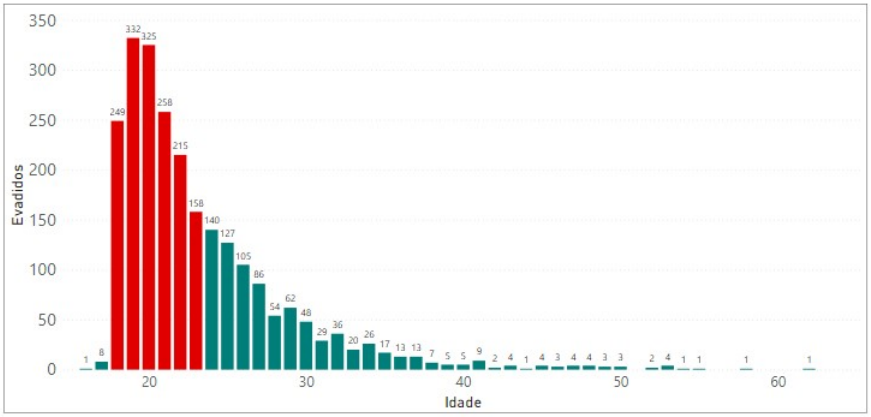
Figura 6 – Total de evasões em relação ao percentual cursado




Fonte: Autor do trabalho

As evasões predominam entre os discentes de 18 e 23 anos. A Figura 7 apresenta o total de discentes evadidos em relação a faixa etária, destacando-se com a cor vermelha a faixa de idade predominante. Para melhor entendimento, é interessante a comparação entre os dados da Figura 3 que apresentam os ingressos por idade, além das análises relativas à permanência dos discentes, presentes na Figura 5 e Figura 6.

Figura 7 – Total de evasões em relação à idade




Fonte: Autor do trabalho

## 1.2 Objetivos

Este trabalho tem como objetivo obter uma caracterização do problema de evasão no ICEA, mediante técnicas mineração de dados nas bases fornecidas pela instituição. A seguir, são definidos os objetivos específicos, para alcançar o objetivo geral desta pesquisa:

- Analisar as características dos discentes evadidos e dos que tendem a evadir e descrever as especificidades de cada grupo.
- Analisar quais são exemplos anômalos entre os discentes evadidos.
- Avaliar e apresentar o conhecimento gerado com os resultados obtidos.

## 1.3 Justificativa

Em seu trabalho, DIAS et al. (2010) apontam a evasão como um dos "males" que afligem as instituições de ensino, comentam que os índices de evasão nas instituições realmente são altos e que vem sendo uma realidade cada vez mais presentes nas IES. Para eles, o fato de um discente abandonar um curso reflete em desperdícios sociais, acadêmicos e econômicos. Nas IES privadas, representa diminuição na receita, já nas IES públicas são gastos de recursos públicos investidos com a ausência do retorno esperado.

Segundo Silva Filho et al. (2007), a problemática da evasão universitária é vista como um dos fatores mais preocupantes nas IES de modo geral. Ao observar os resultados apontados em análises quantitativas em dados disponibilizados pela Seção de Ensino do ICEA, compreende-se que o instituto engloba-se neste quadro, onde há um alto índice de evasão. Infere-se assim, que a instituição vem enfrentando também este problema e há uma grande necessidade do entendimento a cerda do assunto e das possíveis causas ou fatores influenciadores, além da aplicação de medidas corretivas ou protetivas para redução desta taxa.

Com a pesquisa do autor Bispo (1998), percebe-se o quão é importante o uso de ferramenta de auxílio à tomada de decisões. Segundo Bispo (1998), atualmente, os administradores de organizações em geral têm como uma de suas principais funções, o levantamento e análise de informações que os levaram às decisões necessárias para o gerenciamento. O autor ressalta o quão fundamental e indispensável é a utilização de um suporte computacional para levantamento e análise nas informações do negócio. Ainda alega que todas informações tomadas atualmente deverão ser baseadas em dados e/ou fatos comprovados.

Este cenário justifica o presente trabalho, que propõe a aplicação de técnicas de mineração de dados para a caracterização do problema de evasão. Para a escolha destas técnicas, foram levados em consideração, a importância do uso apontada por Bispo (1998) e casos de sucesso em problemas similares como o trabalho de Gonçalves et al. (2018).

## 1.4 Metodologia

Esta seção é voltada à descrição da metodologia aplicada para o desenvolvimento deste trabalho. Pode-se resumir as etapas deste trabalho em: entendimento do problema, seleção e preparo dos dados, aplicação dos algoritmos, e por fim, experimentos e análise dos resultados. A metodologia aplicada nesta pesquisa é baseada no processo *Knowledge Discovery in Databases* (KDD). Este processo possui um fluxo de execução já definido e muito utilizado no campo da mineração de dados.

A pesquisa iniciará com o entendimento do problema, após, será realizada a revisão bibliográfica sobre artigos e trabalhos similares e relacionados ao tema, que norteará o desenvolvimento deste trabalho.

Após entendimento do problema, será dado início a seleção dos dados. As solicitações necessárias serão realizadas junto à Seção de Ensino. Todos os dados obtidos serão analisados, pré-processados e anonimizados. Além da execução de transformações e enriquecimentos a partir dos dados iniciais.

Em seguida, iniciará a escolha dos algoritmos e tarefas de *data mining*, levando em consideração os requisitos e objetivos, o problema e o conjunto de dados disponível. Com os resultados obtidos através dos algoritmos, será realizada a análise e escrita dos resultados.

## 1.5 Organização do trabalho

Este trabalho está estruturado em 4 capítulos. O presente capítulo contém a contextualização e análises iniciais do problema com maior nível detalhes, objetivos e justificativa. O Capítulo 2 apresentará a revisão bibliográfica, abordando em detalhes definições específicas tratadas na pesquisa e trabalhos relacionados ou similares a este. O Capítulo 3 descreve o desenvolvimento, abordando em detalhes as etapas de coleta, pré-processamento, transformação e enriquecimento dos dados. Também será descrito todo processo de preparo e aplicação dos algoritmos de mineração de dados e interpretações dos resultados obtidos. E, por fim, o Capítulo 4 contém as considerações finais e sugestões de possíveis trabalhos futuros.

# 2 Revisão bibliográfica

Este capítulo apresenta a fundamentação teórica, buscando conceituar e trazer definições de termos pertinentes e relacionados a esta pesquisa, tais como: evasão universitária, mineração de dados e uma abordagem sobre a linguagem *Python* para análise de dados. O capítulo apresenta também trabalhos relacionados à temática abordada e desenvolvimentos de pesquisas similares.

## 2.1 Evasão universitária

Entende-se como evasão qualquer forma onde o discente saia do seu curso diferente da diplomação, segundo palavras de (dos Santos, 1999). Os tipos de evasão são detalhados com maior clareza no trabalho de DIAS et al. (2010), onde o autor classifica a evasão em três tipos:

- Evasão do curso: onde o discente desliga-se do curso matriculado sem sua diplomação, ocorrendo por transferência interna ou aprovação em outro curso em sua própria instituição;
- Evasão de instituição, onde há o abandono da instituição e o aluno realiza transferência externa ou é aprovado em outra IES;
- Evasão de sistema, onde o discente desliga-se temporariamente ou permanentemente de sua graduação.

As causas da evasão são divididas por DIAS et al. (2010) em causas internas, onde há problemas relacionados à infraestrutura da IES, corpo docente e assistência socioeconômica, e causas externas, ligadas aos fatores como a falha do discente na escolha do curso, dificuldades escolares, descontentamento com curso e a futura profissão, razões socioeconômicas e problemas pessoais.

## 2.2 Trabalhos relacionados

Existem várias pesquisas direcionadas a temática da evasão universitária, inclusive realizadas na UFOP. Outros trabalhos apresentam características ou objetivos similares ao propósito desta pesquisa. Hoje, qualquer trabalho que envolve a utilização de técnicas de mineração de dados educacionais é classificado como EDM. A seguir, será apresentado trabalhos similares ao tema desta pesquisa, classificados ou não como EDM.

Do Carmo (2018) buscou compreender o problema da Evasão na Universidade Federal de Viçosa (UFV). A pesquisa procurou identificar os evadidos na instituição entre os anos de 2015 e 2016, que representa 36,96%. Constatou-se que os fatores que levam à evasão vão além de questões socioeconômicas, estando muitas vezes ligados ao funcionamento da IES e ao contexto de vida dos discentes analisados. O trabalho possui como características similares a esta pesquisa o fato da identificação dos motivos da evasão, porém com o uso de metodologias diferentes ao deste trabalho.

A seguir temos algumas pesquisas desenvolvidas na UFOP relacionadas a evasão. Passos (2016) apresenta um estudo sobre a evasão e diplomação no curso de Administração da universidade. O trabalho teve como objetivo identificar os fatores ligados a percepção e motivos acerca da escolha do curso, relacionando estes fatores a probabilidade de evasão. Outra pesquisa realizada por Castro and Gouvêa (2014), diz respeito a um estudo, buscando detectar os fatores que influenciam a evasão/retenção dos discentes do Instituto de Ciências Exatas e Biológicas (ICEB), utilizando técnicas paramétricas em análise de sobrevivência. Essas pesquisas citada foram realizadas na UFOP, abordando a temática da evasão, mas não são classificadas como EDM. Utilizaram métodos estatísticos ou análises como pesquisas pessoais para obtenção dos resultados finais.

Gonçalves et al. (2018) apresentam um trabalho similar a este trabalho. Sua pesquisa apresenta um caso de sucesso com a aplicação de mineração de dados buscando obtenção de conhecimento útil no Instituto Federal do Maranhão (IFMA). O trabalho utilizou um método de classificação onde tornou-se possível predizer se um novo discente está propenso a evadir. Com os resultados, o autor indica que é possível tomar medidas com a finalidade de reduzir a taxa de evasão. Este trabalho apresenta uma pesquisa voltada ao entendimento da evasão, com utilização de técnicas de mineração de dados, sendo classificado como EDM, porém foco é voltado à predição.

## 2.3 O processo de descoberta de conhecimento

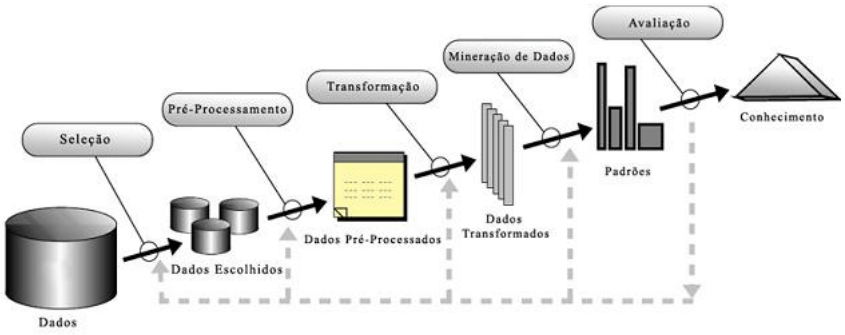
Um termo associado ao tema de mineração de dados é o KDD. A tradução do termo em português consiste em descoberta de conhecimento em banco de dados e pode ser entendido como um processo de extração de dados que visa a obtenção de informação que não é óbvia, anteriormente desconhecida e com potencial de ser útil.

Camilo and Silva (2009) abordam a associação entre KDD e mineração de dados, ressaltando que o entendimento da relação entre KDD e mineração de dados varia de autor para autor. Segundo eles, alguns autores consideram que ambas são práticas sinônimas, já outros comentam que o KDD consiste no processo da extração de conhecimento, já mineração de dados é apenas uma das atividades deste processo.

Calil et al. (2008) comentam que o KDD é uma alternativa eficiente para transformação de dados coletados em conhecimento e apresenta em seu trabalho as etapas deste processo:

- Seleção de Dados: prevê a coleta e seleção dos dados;
- Limpeza: prevê a análise dos dados coletados, verificando a existência de ruídos, tratamento de valores ausentes, entre outras;
- Transformação ou Enriquecimento dos Dados: dedica-se à incorporação e criação de novos dados a partir dos já existentes;
- Mineração de Dados: consiste na aplicação de um algoritmo que, efetivamente, procura por padrões/relações e regularidades, em um determinado conjunto de dados;
- Interpretação e Avaliação: verifica a qualidade do conhecimento (padrões) descoberto, procurando identificar se o mesmo auxilia a resolução do problema original que motivou a realização do processo KDD.

Figura 8 – Uma visão geral das etapas do processo KDD




Fonte: Fayyad et al. (1996)

A Figura 8 apresenta o fluxograma com as etapas do KDD. O processo inicia-se com dados, objetivando a obtenção de conhecimento.

Mineração de dados resume-se em um processo de extrair conhecimento de grandes volumes de dados (De Amo, 2004). Segundo Camilo and Silva (2009) o termo mineração de dados, do inglês *data mining*, surge no final da década de 80, onde tinha-se um panorama caracterizado pela expressão "empresas eram ricas em dados e pobres em informação".

Braga (2005) reforça que o processo de mineração de dados consiste em ferramentas e métodos automáticos para identificação de padrões em conjuntos dados já coletados,

ausentes de tendências e limitações de uma análise humana. Ainda define de maneira genérica, que mineração de dados é um conjunto de técnicas para descrever e predizer informações em grandes conjuntos de dados, além de afirmar que as técnicas de mineração estão inseridas em um processo de descoberta de conhecimento.

O processo de mineração de dados é composto por várias tarefas. Para Camilo and Silva (2009) as tarefas mais comuns são: descrição, classificação, estimação, predição, agrupamento e associação. De Amo (2004) comenta que as técnicas de mineração de dados surgiram quando as empresas passaram a observar a grande quantidade de dados que as empresas possuíam armazenados. Segundo o autor, inicialmente as tarefas de mineração consistiam apenas em extrair informações das massivas bases de dados de forma mais automática possível e que atualmente, o conceito está associado com a obtenção de conhecimento significativo sobre os dados extraídos.

De Amo (2004) ressalta o quão é importante distinguir o que são as técnicas e as tarefas de mineração de dados. Para ele, as técnicas consistem na especificação de métodos que nos garantam como descobrir os padrões que nos interessam. O autor cita alguns exemplos de técnicas de mineração, entre elas estão: técnicas estatísticas, técnicas de aprendizado de máquina e técnicas baseadas em crescimento poda-validação. Já a tarefa, consiste na especificação do que estamos querendo buscar nos dados, que tipo de regularidades ou categoria de padrões temos interesse em encontrar, ou que tipo de padrões poderiam nos surpreender. A seguir, temos a descrição sucinta das tarefas utilizadas neste trabalho, baseado no estudo de De Amo (2004):

- Análise de *clusters*: Diferentemente da classificação e predição onde os dados de treinamento estão devidamente classificados e as etiquetas das classes são conhecidas, a análise de *clusters* trabalha sobre dados onde as etiquetas das classes não estão definidas. A tarefa consiste em identificar agrupamentos de objetos, agrupamentos estes que identificam uma classe;
- Análise de *outliers*: os dados que fogem do padrão normal do banco de dados são mais interessantes que as classes predominantes. Conhecidos como *Outliers* ou anomalias, esta tarefa é de extrema importância quando a necessidade é examinar características anormais.

Segundo Witten et al. (2017), o aprendizado supervisionado como o próprio nome diz, tem um supervisor externo, ou seja, é conhecido a saída desejada para cada exemplo. As tarefas de classificação e regressão são exemplos de aprendizado supervisionados. Os algoritmos não-supervisionados, por outro lado, permite abordar problemas com pouca ou nenhuma ideia do que os resultados podem apresentar. Pode-se derivar estrutura de dados, onde não necessariamente sabe-se o efeito das variáveis. As tarefas de clusterização e detecção de *outliers* são exemplos de aprendizado não-supervisionados.

## 2.4 Utilização da linguagem *python* em mineração de dados

Coelho (2007) define a linguagem *Python* como uma linguagem de programação dinâmica e orientada a objetos, podendo ser utilizada em qualquer tipo de aplicação, científica ou não. Comenta também que, possui uma sintaxe simples e clara, suporte à interação com outras linguagens e possuir uma vasta biblioteca padrão. O autor reforça também que a linguagem é frequentemente associada com grandes ganhos em produtividade, produção de programas de alta qualidade e de simples manutenção. Segundo o autor, a linguagem começou a ser desenvolvida no final dos anos 80, na Holanda, por Guido van Rossum, que até os dias de hoje continua liderando o desenvolvimento da evolução da linguagem.

Dentre várias características da linhagem, Coelho (2007) destaca como cruciais os seguintes fatores: multiplataforma, portabilidade, *software* livre, extensibilidade, orientação a objetos, tipagem automática e forte, código legível, flexibilidade, operação com arquivos e uso interativo.

Fernandes (2019) aponta a programação na linguagem *Python* como uma das linguagens mais populares para a computação científica. Para utilização do *Python* voltado às análises de dados, necessita-se de ferramentas, denominadas como *Application Programming Interface* (API) ou simplesmente como bibliotecas. Uma definição contida no *site* Mundo Digital [1] define de forma clara esta terminologia. Segundo esta definição, uma API consiste em uma interface de programação de aplicação, e possuem um conjunto de rotinas e padrões estabelecidos por um *software* para a utilização das suas funcionalidades por aplicativos que não pretendem envolver-se em detalhes da implementação do software, mas apenas usar seus serviços. A seguir contém algumas destas bibliotecas, segundo definições constadas no trabalho de Fernandes (2019):

- *Scikit-learn*: é uma ferramenta que tem sido fundamental para permitir que o *Python* seja uma linguagem de programação de uma ciência de dados produtiva. A autora reforça que a *scikit-learn* expõe uma ampla variedade de algoritmos de aprendizado de máquina, supervisionados e não supervisionados, usando uma interface consistente e orientada a tarefas, permitindo uma comparação fácil de métodos para uma determinada aplicação; foi utilizado o algoritmo de clusterização *K-Means*[2] e o algoritmo de detecção de *outliers NearestNeighbors*[3], pertencentes à esta biblioteca;

- *Pandas*: é uma biblioteca *open source*, que fornece estruturas de dados de alto desempenho e fáceis de usar com ferramentas de análise de dados para a linguagem de

[1] https://geoftp.ibge.gov.br/organizacao_do_territorio/estrutura_territorial/localidades/Google_KML/

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

[3] https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.NearestNeighbors.html

programação *Python*. Fornece estruturas e funções de dados de alto nível projetadas para tornar a manipulação de dados estruturados ou tabulares rápidos, fáceis e expressivos;

- *Matplotlib*: é uma biblioteca para criar gráficos bidimensionais de matrizes em *Python* e pode ser usada sem prejuízos no paradigma orientado a objetos.

# 3 Desenvolvimento

Nesta seção encontra-se o detalhamento das etapas desenvolvidas, seguindo o processo KDD.

## 3.1 Seleção dos dados

Nesta seção, busca-se entender as necessidades das partes interessadas e o cenário em questão. Entender o problema, estudar os tópicos e pontos, além de pesquisar trabalhos e casos de sucesso similares foi a primeira concluída.

Junto à seção de ensino do instituto, foram verificado as fontes existentes e quais dados estas continham em relação ao problema trabalhado. Das fontes de dados apresentadas, três foram selecionadas para o desenvolvimento das análises. Elas foram disponibilizados em formato de planilhas. Por se tratar de dados sensíveis, contendo informações pessoais de cada aluno, houve a necessidade de realizar anonimização destes dados. A seguir temos a descrição das bases:

- Dados gerais: contém várias informações gerais de todos discentes ingressados na instituição por matrícula.
- Notas dos discentes: a fonte apresenta a nota de cada disciplina feita por cada aluno.
- Dados dos discentes evadidos: contém informações específicas dos discentes que evadiram e que não estão presentes na fonte de dados gerais.

Nas Tabelas 1, 2 e 3 a primeira coluna refere-se ao atributo de dados e a segunda coluna contém a descrição. Estes atributos permanecem, após a limpeza de dados. Os dados de matrícula foram anonimizado, por se tratar de um dado sensível.

Tabela 1 – Descrição da planilha de dados gerais

| **Dado** | **Descrição** |
|---|---|
| Matrícula | Matrícula do aluno, ao qual foi anonimizado |
| Sexo | Sexo do aluno |
| Cidade de nascimento | Cidade de nascimento do aluno |
| Estado de nascimento | Estado de nascimento do aluno |
| País de nascimento | País de nascimento do aluno |
| Bairro familiar | Bairro onde reside a família do aluno |
| Cidade familiar | Cidade onde reside a família do aluno |

| **Dado** | **Descrição** |
|---|---|
| Estado familiar | Estado onde reside a família do aluno |
| País familiar | País onde reside a família do aluno |
| Bairro acadêmico | Bairro onde o aluno reside, podendo ser igual ou não ao bairro familiar |
| Cidade acadêmico | Cidade onde o aluno reside, podendo ser igual ou não ao bairro familiar |
| Estado acadêmico | Estado onde o aluno reside, podendo ser igual ou não ao bairro familiar |
| País acadêmico | País onde o aluno reside, podendo ser igual ou não ao bairro familiar |
| República | República onde o aluno está residindo |
| Tipo de república | Tipo de república que o aluno reside |
| Data de nascimento | Data de nascimento do aluno |
| Ano de admissão | Ano de ingresso no instituto |
| Semestre de admissão | Semestre de ingresso no instituto |
| Descrição do modo de admissão | Descrição do modo de admissão |
| Pontuação no vestibular | Pontuação do vestibular/ENEM |
| Código do curso de admissão | Curso que o aluno ingressou, podendo alterar |
| Origem | Origem dos discentes, aplicado somente aos casos quem que os discentes vieram por transferência externa de outra faculdade |
| Turno | Turno regular de aulas |
| Descrição da situação do aluno | Descrição da situação em que o aluno se encontra |
| Carga horária do curso | Carga horária do curso matriculado |
| Carga horária cursada | Carga horária cursada pelo aluno |
| Código do curso atual | Código do curso atual que o aluno está matriculado |
| Curso | Curso atual que o aluno está matriculado |
| Código de habilitação | Apresenta o código de habilitação |
| Código de enfase | Apresenta o código de ênfase |
| Data de ingresso | Data de ingresso do aluno no instituto |
| Participou de políticas afirmativa | Informa se o aluno participou das políticas afirmativas |
| Usou política afirmativa | Informa se o aluno usou políticas afirmativas. |
| Mobilidade de concorrência | Informa a modalidade de concorrência |

| **Dado** | **Descrição** |
|---|---|
| Mobilidade de concorrência homologada | Informa se houve homologação da modalidade de concorrência |
| Ano de diplomação | Informa o ano que o aluno diplomou, contendo dados apenas dos discentes formados. Os demais estão representados como nulo. |
| Semestre de diplomação | Informa o semestre que o aluno diplomou, contendo dados apenas dos discentes formados. Os demais estão representados como nulo. |

Fonte: Autor do trabalho

Tabela 2 – Descrição da planilha de notas

| **Dado** | **Descrição** |
|---|---|
| Ano | Referente ao ano de aplicação da disciplina |
| Semestre | Referente ao semestre de aplicação da disciplina |
| Código da disciplina | Código da disciplina |
| Descrição | Nome da disciplina |
| Código da turma | Código da turma de determinada disciplina |
| Matrícula | Matrícula do aluno |
| Cor da pele | Cor da pele do aluno |
| Sexo<br>Excluído | Sexo do aluno |
| Código do curso | Curso do aluno |
| Caráter | Caráter da disciplina |
| Média final | Pontuação final do aluno na disciplina |
| Exame especial | Pontuação do aluno no exame especial, se necessário |
| Faltas | Total de faltas do aluno na disciplina |
| Situação | Descrição da situação final do aluno na disciplina |
| Gravação | Data de lançamento no sistema no sistema |
| Tipo de escola | Tipo de escola que o aluno frequentou no primeiro e segundo grau |

Fonte: Autor do trabalho

Tabela 3 – Descrição da planilha de evadidos

| **Dado** | **Descrição** |
|---|---|
| Ano de evasão | Ano de evasão do aluno |
| Semestre de evasão | Semestre de evasão do aluno |
| Matrícula | Matrícula do aluno |
| Descrição | Descrição geral do motivo de evasão do aluno |
| Destino | Aplicado aos casos de transferência externa, descreve para qual universidade o aluno se transferiu |

Fonte: Autor do trabalho

Por fim, outra base de dados necessária para enriquecimento dos dados foi a base do Instituto Brasileiro de Geografia e Estatística (IBGE) [1]. Nesta base temos a latitude e longitude de todos as cidades brasileiras.

## 3.2 Limpeza e pré-processamento

Estas etapas são realizadas utilizando a ferramenta da *Microsoft* chamada *Power Query*, presente no *Software Power Bi*. Iniciam-se com a anonimização de todos os dados sensíveis seguindo as normas da Lei Geral de Proteção de Dados Pessoais (LGPD)[2]. Para cada fonte de dados, foram pontuados os dados pessoais ou sensíveis, são eles:

- Dados gerais: matrícula do aluno, nome do aluno, Comprovante de Situação Cadastral (CPF) do aluno e *e-mail* do aluno;
- Notas dos discentes: matrícula do aluno, nome do professor;
- Dados dos discentes evadidos: matrícula do aluno, nome do aluno e *e-mail* do aluno.

Os dados de matrícula foram anonimizados, de tal forma que, para cada matrícula existente nas três bases de dados, foi gerado um identificador único para cada uma. Em seguida, houve a exclusão do campo matrícula de todas as fontes. Este identificador foi escolhido para realização dos relacionamento entre as fontes de dados, sendo utilizado como chave primária. As demais informações foram excluídas por se tratar de informações não relevantes e desnecessárias. Com esta etapa concluída, foi possível ter uma fonte de dados sem a presença de dados que venham infringir a LGPD, colocando exposto dados pessoas e sensíveis dos discentes do instituto.

[1] https://geoftp.ibge.gov.br/organizacao_do_territorio/estrutura_territorial/localidades/Google_KML/

[2] http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm

Dando sequência no pré-processamento, todas as colunas de dados que foram consideradas desnecessárias, replicadas ou sem relevância foram excluídas. As demais colunas foram convertidas em relação ao tipo de dados.

## 3.3 Transformação e enriquecimento dos dados

Através do *Power Query*, foram criadas novas colunas que permitiram o enriquecimento dos dados. São elas:

- Carga horária cursada: relação entre a carga horária que o aluno cursou e total de horas do curso que ele está matriculado, obtendo o percentual cursado;
- Checagem de transferência: analisa se o aluno mudou de curso dentro do instituto; se o curso de admissão for igual ao curso de atual, temos o estado de permanência no curso, caso contrário, temos o estado de transferência interna;
- Idade de evasão: calcula a idade que o aluno evadiu.
- Idade de ingresso: calcula a idade em que o aluno ingressou no ICEA.
- Semestres de permanência: calcula quantos semestres os discentes evadidos permaneceram na instituição, através da diferença entre a data de evasão e data de ingresso;

Em notas dos discentes foram criadas as seguintes colunas:

- Resultado final: cria-se uma coluna de resultado final, analisando se um aluno foi submetido ao exame final; se o aluno não realizou exame final, o resultado será igual ao campo média final, caso contrário, será igual ao resultado do exame especial, desta maneira temos a mesclagem da coluna média final e exame final;
- Total disciplina: calcula o total de disciplina que determinado aluno se matriculou;
- Total aprovado: calcula o total de disciplina aprovada de cada aluno;
- Total reprovado: calcula o total de disciplina reprovada de cada aluno;
- Total trancado: calcula o total de disciplina trancada de cada aluno;
- Total cancelado: calcula o total de disciplina cancelada de cada aluno;
- Total de exames finais necessários: calcula o total de exames finais que cada aluno necessitou;

- Desistência de exame final: calcula quantos exames finais o aluno não realizou, mesmo tendo a possibilidade;
- Coeficiente: média das notas de cada aluno;
- Percentual de aprovação: relação entre o total aprovado e o total de disciplinas feitas por cada aluno.
- Percentual de reprovação: relação entre o total reprovado e o total de disciplinas feitas por cada aluno.
- Percentual de trancado: relação entre o total trancado e o total de disciplinas feitas por cada aluno.
- Percentual de cancelado: relação entre o total cancelado e o total de disciplinas feitas por cada aluno.
- Percentual de necessidade de exame final: relação entre o total de exames finais necessários e o total de disciplinas feitas por cada aluno.
- Percentual de desistência em exames finais: relação entre o total de desistência de exame final e o total de total de exames finais necessários.

Outro atributo criado refere-se ao cálculo das distâncias em linha reta das cidades familiar de cada aluno a cidade de João Monlevade, utilizando as coordenadas geográficas, obtendo assim uma referência numérica.

A fonte dados dos discentes evadidos foi mesclada à fonte de dados gerais, tendo como chave o identificador gerado através da matrícula. Também foi mesclado o cálculo da distância.

### 3.3.1 *Views* de atributos

Uma *View*, ou visão, é uma forma em que podemos analisar a base de dados através de consultas. A criação desta visualização facilita o trabalho através da sumarização de atributos. Nesta seção são definidos quais atributos irão compor determinada *View*. A seguir, estas são apresentadas:

- *View* 1: reúne dados específicos referentes aos discentes evadidos; cada exemplo descreve um aluno evadido. Esta *view* contém 2286 exemplos e 38 atributos, sendo eles: sexo, ano e semestre de evasão, cor da pele, tipo de escola frequentada pelo aluno nos ensinos básicos, turno, análise de transferência interna, modo de admissão, uso e participação em políticas afirmativas, modalidade de concorrência e estado de homologação, curso de admissão e curso atual, idade de ingresso e evasão, faculdade

de origem e destino aos casos cabíveis, motivo de evasão apontado, descrição da evasão, motivo de evasão, semestre de permanência no ICEA, carga horária cursada, percentual cursado referente ao curso matriculado, coeficiente, cidade de residência do aluno, endereço da família do discentes contendo cidade, estado e país, distância da cidade da família à cidade de João Monlevade, análise referente se o aluno morava na mesma cidade da família. Também foram analisados os dados do total de disciplinas matriculadas, percentual de aprovações, reprovações, trancamento e cancelamento, além do percentual do total exames finais necessários em relação ao total de disciplinas e o percentual de exames com desistência.

- *View* 2: reúne dados gerais dos discentes, independente de sua situação (evadido, diplomado, matriculado, trancado ou cancelado) e informações históricas dos discentes; cada exemplo descreve um aluno da universidade; esta *view* é composta por 4767 exemplos e 30 atributos, sendo eles: sexo, cor da pele, sexo, idade de ingresso, descrição do modo de admissão, período do ano em que o aluno ingressou, turno de ingresso, tipo de escola que o aluno frequentou, análise do uso e participação de políticas afirmativas, mobilidade de concorrência, país, estado e cidade onde vive a família do aluno, cidade onde o aluno reside, checagem se o aluno reside ou não com a família, distância entre a cidade onde reside a família à João Monlevade e o curso que o aluno foi admitido, curso atual do aluno, sendo que este pode haver variação, análise de transferência interna, apresentando se houve ouve ou não transferência interna, ano e semestre de admissão, percentual cursado. Também foram inseridos dados de rendimentos dos discentes como coeficiente, o total de disciplinas matriculadas, percentual de aprovação, reprovação, trancamento e cancelamento. Também o percentual da necessidade de exame finais e o percentual de desistência nos exames finais.

- *View* 3: reúne informações do desempenho dos discentes; cada exemplo corresponde a um aluno. Está *view* é composta por 4663 exemplos e 8 atributos, sendo eles: coeficiente, total de disciplinas, percentual cursado, aprovado, reprovado, trancado, cancelado e com necessidade de exames finais.

## 3.4 Mineração de dados

Esta seção contém a descrição da transformação de dados simbólicos e normalização, a criação das matrizes de correlação e a parametrização e aplicação dos algoritmos.

### 3.4.1 Transformação dos dados simbólicos e normalização

A transformação dos dados simbólicos em dados numéricos consiste na conversão de todos atributos da *view* do tipo texto em atributos do tipo inteiro. Para cada elemento do conjunto de atributos do tipo texto, é gerado um número inteiro único, substituindo a cadeia de caractere por este valor. Neste trabalho, esta conversão foi realizada com o uso de um recurso presente na biblioteca do *sklearn* [3], conhecido como *preprocessing*. Após a execução deste método, todos os atributos da *view* são do tipo inteiro ou decimal, característica fundamental para que seja possível realizar a normalização dos dados.

Normalizar os dados é essencial para aplicação dos algoritmos de *data mining*. Através da técnica de normalização, temos como resultado a padronização dos valores da *view*, de tal forma que todos os atributos de dados estarão na mesma escala, impedindo que algum valor predomine sobre outro ou gere algum tipo de ponderação indesejada nos resultados. Neste trabalho, foi aplicada a técnica de normalização por reescala em todas as *views* geradas. A normalização de cada atributo j, de um elemento $x_i$ pode ser calculada como:

$$x_{norm.(i,j)} = \frac{x_{i,j} - min_j}{max_j - min_j}$$

Estas etapas são fundamentais para execução dos algoritmos de agrupamento e análise de *Outliers*.

### 3.4.2 Escolha, preparo e aplicação dos algoritmos

Neste trabalho, houve a aplicação de duas tarefas de mineração de dados, sendo elas: análise de *clusters* e análise de *outliers*.

O algoritmo *K-Means* tem como objetivo encontrar similaridades, independente da situação (evadido, diplomado, matriculado, cancelado e trancado) do aluno. O algoritmo agrupará os dados conforme o número de *cluster* passado pelo argumento k. Os exemplos são agrupados de acordo com a média da distância de cada ponto até o centroide.

Fundamental na utilização dos algoritmos de agrupamento é a escolha do total de grupos. A quantidade desejada de grupos é um dos parâmetros de entrada do algoritmos, representado pela variável k. O método *Elbow* auxilia na determinação do melhor número de *clusters*. No método *Elbow*, o melhor número para k é definido no momento onde não há variação significante em relação ao aumento do número de *clusters*.

As Figuras 9a, 9b e 9c apresentam os resultados da aplicação do método para cada *view*. A Tabela 4 apresenta os números de *clusters* escolhidos para cada *view*.

[3] https://scikit-learn.org/stable/

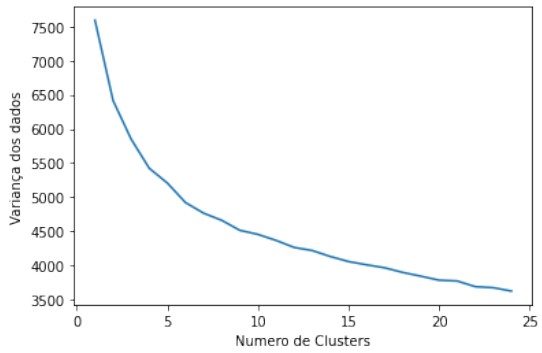
A tarefa de agrupamento foi aplicada em todas as *views* geradas. Os resultados são utilizados na etapa de avaliação e interpretação dos resultados. Após a identificação dos grupos em cada *view* gerada, é possível predizer em qual grupo um novo exemplo pertencerá, baseando-se na similaridade dos seus dados e os exemplos já agrupados, utilizando o método *predict* da biblioteca *scikit-learn*. Mais informações sobre o algoritmo *K-means* e a funcionalidade da predição podem ser encontradas no Anexo A.4.

Segundo Varella (2008), a técnica estatística PCA, também conhecida como análise dos componentes principais, tem como objetivo a redução da dimensionalidade do sistema, cujos componentes principais são obtidos a partir da matriz de covariância dos atributos originais. Neste trabalho, todos as *views* foram reduzidos em três componentes, facilitando a representação dos resultado em uma plotagem de um gráfico tridimensional.

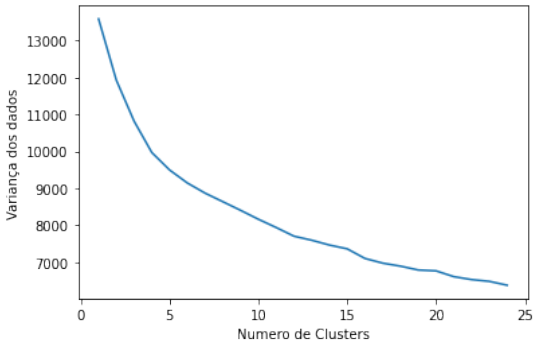
Após a criação dos *clusters*, parte-se para outra etapa: encontrar os exemplos mais anômalos nas *views*. Para este caso, temos a tarefa de detecção de *outliers*. O algoritmo *NearestNeighbors* de forma resumida, calcula as distâncias entre os exemplos e se comporta como um algoritmo não-supervisionado, devido a ausência de rótulos descritivos.

A Seção 3.5 apresenta as análises dos resultados obtidos pelos algoritmos apresentados nesta seção.

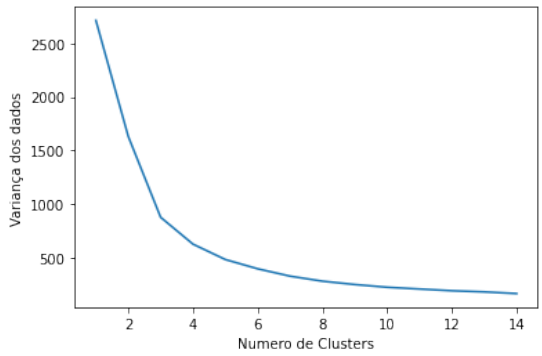
Figura 9 – Análise do método *Elbow* para as *views*




(a) *View* 01




(b) *View* 02




(c) *View* 03

Tabela 4 – Melhor número de *clusters* para cada *view*

| ***View*** | **Número de *Clusters*** |
|---|---|
| *View* 01 | 5 |
| *View* 02 | 6 |
| *View* 03 | 5 |

Fonte: Autor do trabalho

## 3.5 Análise dos resultados

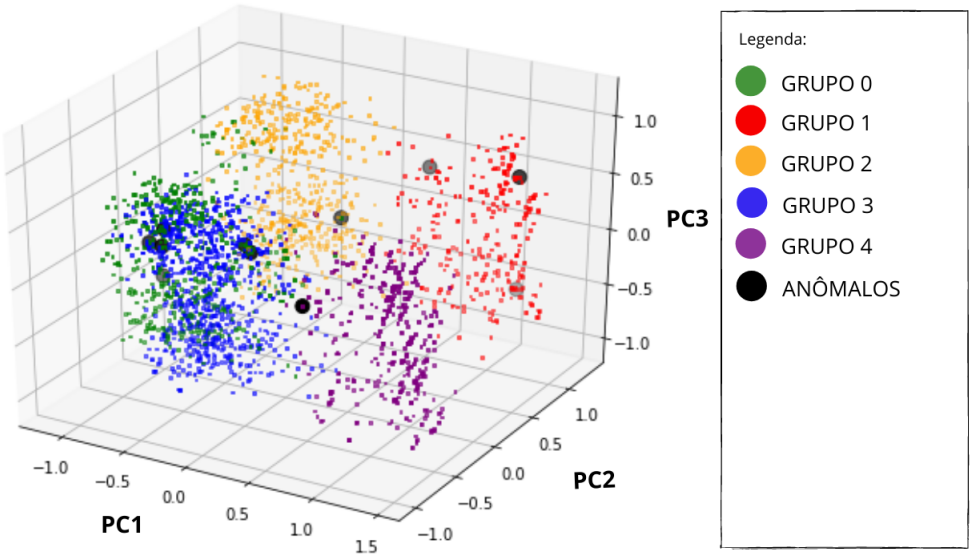
Esta seção é voltada às análises realizadas através das tarefas de mineração de dados. Analisar e avaliar os resultados obtidos é a última etapa do processo KDD.

### 3.5.1 Apresentação dos resultados da *view* 01

A *view* 1 refere-se ao conjunto de discentes evadidos, que servirá de base para contrastarmos com as análises efetuadas nas demais *views*. Com a aplicação da tarefa de agrupamento houve a geração de cinco grupos de acordo com os resultados do método de *Elbow* e algoritmo *kmeans*. Cada grupo, reúne discentes evadidos com motivos, rendimentos ou características parecida.

A Figura 10 apresenta os grupos gerados com a redução da dimensionalidade obtida utilizando o método PCA. As cores representam os exemplos contidos em cada grupo.

Figura 10 – Representação tridimensional do agrupamento dos dados da *view* 01

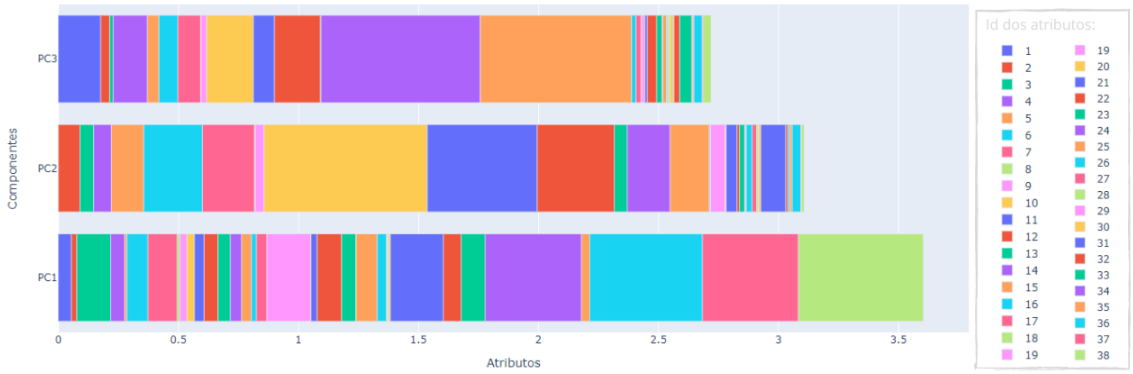



Fonte: Autor do trabalho

O algoritmo *kmeans* foi aplicado na base dados agrupando os exemplos de acordo com suas similaridades. A redução da dimensionalidade foi realizada, reduzindo os atributos para 3 componentes, permitindo a visualização em 3 dimensões.

A Figura 11 apresenta a contribuição dos atributos nos componentes, que representam os eixos PC1, PC2 e PC3 da Figura 10.

Figura 11 – Contribuição dos atributos da *view* 1 em cada componente




Fonte: Autor do trabalho

A Tabela 5 complementa, apresentando a contribuição de cada atributo em cada componente. Os atributos que mais contribuem em cada componente estão enfatizados.

Tabela 5 – Detalhamento da contribuição ($\alpha$) de cada atributo nos componentes PC1, PC2 e PC3 da *view* 1

| **Id** | **Atributo** | $\alpha$ **PC1** | $\alpha$ **PC2** | $\alpha$ **PC3** |
|---|---|---|---|---|
| 1 | SEXO | 5% | 0% | **18%** |
| 2 | ANO EVASAO | 2% | 9% | 4% |
| 3 | SEMESTRE EVASAO | **14%** | 6% | 1% |
| 4 | TURNO | 6% | 7% | **14%** |
| 5 | COR DA PELE | 1% | 13% | 5% |
| 6 | TIPO ESCOLA | 9% | **24%** | 8% |
| 7 | TIPO VESTIBULAR | 12% | **22%** | 10% |
| 8 | CHECK TRANSF INTERNA | 1% | 0% | 0% |
| 9 | MODO ADMISSAO | 3% | 4% | 2% |
| 10 | PARTICIPOU POLITICA | 3% | **68%** | **19%** |
| 11 | USOU POL AFIRMATIVA | 4% | **46%** | 9% |
| 12 | MODALIDADE CONCORRENCIA | 6% | **32%** | **19%** |
| 13 | MODALIDADE HOMOLOGADA | 5% | 5% | 0% |

| **Id** | **Atributo** | $\alpha$ **PC1** | $\alpha$ **PC2** | $\alpha$ **PC3** |
|---|---|---|---|---|
| 14 | COD CURSO ADMISSAO | 5% | **18%** | **66%** |
| 15 | COD CURSO ATUAL | 4% | 16% | **63%** |
| 16 | IDADE INGRESSO | 2% | 0% | 2% |
| 17 | IDADE EVASAO | 4% | 0% | 2% |
| 18 | ORIGEM | 0% | 0% | 0% |
| 19 | DESCRICAO EVASAO | **18%** | 6% | 1% |
| 20 | DESTINO EVASAO | 0% | 0% | 0% |
| 21 | EVASAO PROCESSADO | 3% | 4% | 1% |
| 22 | SEMESTRES DE PERMANENCIA | 10% | 1% | 4% |
| 23 | CARG HORARIA CURSADA | 6% | 2% | 2% |
| 24 | PERC CURSADO | 0% | 0% | 0% |
| 25 | COEFIENTE | 9% | 0% | 2% |
| 26 | CIDADE ACADEMICO | 4% | 2% | 0% |
| 27 | CIDADE FAMILIAR | 0% | 2% | 0% |
| 28 | ESTADO FAMILIAR | 1% | 1% | 1% |
| 29 | PAIS FAMILIAR | 0% | 0% | 0% |
| 30 | DISTANCIA | 1% | 1% | 1% |
| 31 | CHECK ENDERECO | 22% | 10% | 0% |
| 32 | TOTAL DE MATERIA | 7% | 1% | 2% |
| 33 | PERC APROVADO | 10% | 1% | 5% |
| 34 | PERC REPROVADO | **40%** | 0% | 0% |
| 35 | PERC TRANCADO | 3% | 1% | 1% |
| 36 | PERC CANCELADO | **47%** | 4% | 3% |
| 37 | PERC EXAME FINAL | **40%** | 0% | 0% |
| 38 | PERC DESISTENCIA EXAME | 52% | 1% | 3% |

Fonte: Autor do trabalho

A Tabela 6 apresenta valores sobre parâmetros de rendimento dos exemplos dos grupos. Se observarmos os atributos que mais contribuem $\geq 40\%$ ao PC1 e PC2 eles condizem com a média e percentual de aprovação. Também é importante observar o tipo de ingresso na universidade, que tende a indicar a não utilização de politicas afirmativas.

Tabela 6 – Análise dos grupos gerados a partir da *view* 1

| ***Parâmetro*** | **Grupo 0** | **Grupo 1** | **Grupo 2** | **Grupo 3** | **Grupo 4** |
|---|---|---|---|---|---|
| Cor do grupo | Verde | Vermelho | Amarelo | Azul | Roxo |

| ***Parâmetro*** | **Grupo 0** | **Grupo 1** | **Grupo 2** | **Grupo 3** | **Grupo 4** |
|---|---|---|---|---|---|
| Total de exemplos | 465 | 280 | 511 | 725 | 405 |
| Média de coeficiente | 30,36 | 5,62 | 28,22 | 28,81 | 3,33 |
| Percentual cursado médio | 23,34% | 2,14% | 11,51% | 11,34% | 1,42% |
| Percentual de aprovação nas disciplinas | 41,10% | 15,30% | 34,40% | 35,4% | 9,2% |

Fonte: Autor do trabalho

A seguir são elencadas as características descobertas, após as análises das tabelas e figuras anteriores, referentes a cada grupo.

- **Grupo 0**: possui concentração dos exemplos nos *scores* negativos dos componentes PC1 e PC2, com pequena parcela de exemplos no *score* positivo do componente PC2; este grupo, representado pela cor verde na Figura 10, reúne 465 discentes, que mesmo vindo a evadiram, possuem rendimentos significativos em relação aos demais; a média de coeficiente deste grupo é de 30,36, a média em percentual cursado desdes discentes é equivalente a 23,34% e com percentual de aprovação médio nas disciplinas de 41,10%; destaca-se neste grupo a presença de discentes que foram desligados da instituição por atingirem o prazo máximo para diplomação.

- **Grupo 1** possui concentração dos exemplos nos *scores* positivos dos componentes PC1 e PC2; representado pela cor vermelha na Figura 10, reúne 280 discentes que vieram a evadir nos primeiros semestre da graduação; é percebido um baixo rendimento comum neste grupo; a média de coeficiente dos discentes é equivalente a 5,62, cursaram em média 2,14% da graduação e o percentual médio de aprovação nas disciplinas é equivalente a 15,30%.

- **Grupo 2**: possui concentração dos exemplos no *score* negativo do componente PC1 e no *score* positivo do componente PC2; representado pela cor amarela na Figura 10, reúne 511 discentes com bons rendimentos que evadiram nos primeiros anos da graduação; possuem médias de coeficientes e percentual médio de aprovações significativos, mas tem como percentual médio cursado equivalente a 11,51%.

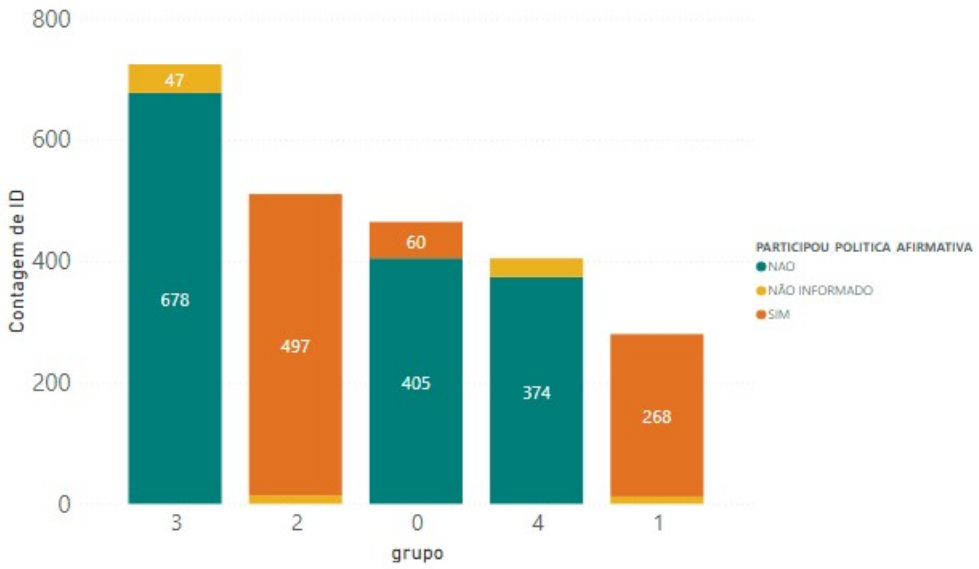
- **Grupo 3**: possui concentração dos exemplos nos *scores* negativos dos componentes PC1 e PC2; representado pela cor azul na Figura 10, reúne 725 discentes com bons rendimentos que evadiram nos primeiros anos da graduação; este grupo possui uma

similaridade com o grupo 2 quando analisa-se a média de coeficientes, percentual cursado e o percentual de aprovação, porém, são diferentes quando observa-se o uso de políticas afirmativas entre os discentes reunidos neles; neste grupo, reúne discentes que não utilizaram as políticas de cotas para ingresso.

- **Grupo 4**: possui concentração dos exemplos no *score* negativo do componente PC2 e no *score* positivo do componente PC1; este grupo, representado pela cor roxa na Figura 10, reúne 405 discentes, que vieram evadir nos primeiros anos da graduação; este grupo possuem discentes com rendimentos similares aos do grupo 2, mas diferenciam quando observa-se a utilização das politicas afirmativas.

A Figura 12 elucida o atributo do uso de politicas afirmativas entre os grupos formados. Em alguns casos, os grupos possuem rendimentos similares, mas o que distingue um grupo de outro é a utilização ou não das politicas afirmativas.

Figura 12 – Uso de políticas afirmativas por grupo



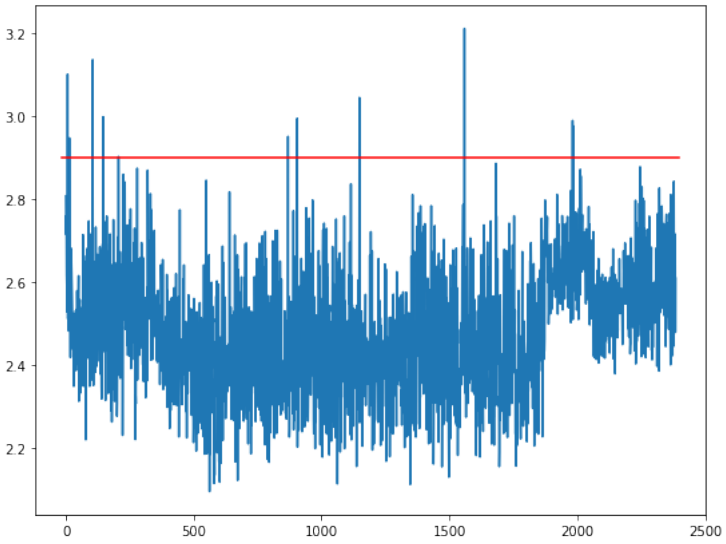

Fonte: Autor do trabalho

A detecção de elementos anômalos é aplicada quando a discrepância é relevante. Observando-se a proposta do trabalho, a detecção de discentes evadidos anômalos é interessante para perceber quais são casos mais distintos do normal. Com o uso desta técnica, houve como principal retorno, o apontamento de discentes que foram desligados por atingirem o prazo máximo, mesmo tendo rendimentos significativamente elevados.

A partir da aplicação da técnica de detecção de *outlies*, foram encontrados 11 exemplos mais anômalos, considerando o ponto de corte de 2.9 das médias das distâncias entre todos os 2384 exemplos, conforme o gráfico da Figura 13. Os exemplos anômalos são

apresentados na Tabela 7. Aproximadamente 45% destes evadiram sem cursar nenhum disciplina. Percebe-se que dois discentes estão com percentual cursado superior a 0, mesmo não cursando disciplinas. Estes, possivelmente, referem-se a erro de entrada ou execução de atividades extra-curriculares. Alguns discentes possuem coeficiente e percentual cursado altos, com baixo número de disciplinas reprovadas, mas mesmo assim vieram a evadir, sendo desligados por prazo máximo.

Figura 13 – Detecção das anomalias na *view* 01




Fonte: Autor do trabalho

Tabela 7 – Exemplos anômalos em relação ao padrão da *view* 01

| **Id** | **Grupo** | **Média** | **Cidade familiar** | **Idade Ingresso** | **Cursado** | **Aprovações** |
|---|---|---|---|---|---|---|
| 50249964 | 4 | 0,0 | Ouro Preto | 24 | 12,5% | 0 |
| 100532940 | 1 | 0,0 | Belo Vale | 20 | 0,0% | 0 |
| 108911010 | 1 | 0,0 | Belo Vale | 20 | 0,0% | 0 |
| 125671875 | 1 | 0,0 | Montalvania | 18 | 0,0% | 0 |
| 50254446 | 0 | 0,0 | Ourinhos | 19 | 7,59% | 0 |
| 92126969 | 0 | 29,9 | Bissau | 25 | 34,00% | 8 |
| 100499184 | 0 | 47,62 | Rio Casca | 28 | 81,72% | 37 |

| **Id** | **Grupo** | **Média** | **Cidade familiar** | **Idade Ingresso** | **Cursado** | **Aprovações** |
|---|---|---|---|---|---|---|
| 58625546 | 0 | 57,18 | Vila Velha | 18 | 104,8% | 59 |
| 75376413 | 0 | 68,4 | Monte Claros | 22 | 101,66% | 55 |
| 41876265 | 0 | 86,96 | São Domingos do Prata | 22 | 104,54% | 46 |
| 25125711 | 0 | 88,72 | Belo Horizonte | 22 | 101,94% | 54 |

Fonte: Autor do trabalho

Assim, a universidade precisa estar atenta aos discentes com bons rendimentos e próximos ao prazo máximo de conclusão, além de elaborar políticas que possam resolver esta situação. Com a aplicação das tarefas de mineração utilizadas neste trabalho, foi possível dividir discentes evadidos em cinco grupos, definir suas características principais e identificar anomalias.

### 3.5.2 Apresentação dos resultados da *view* 02

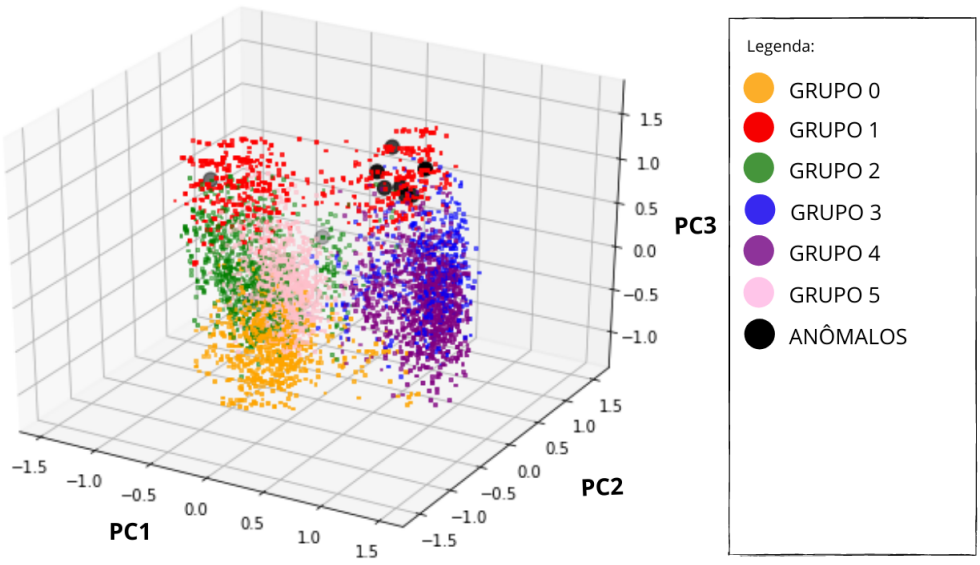
A *view* 2 busca analisar os dados gerais dos discentes, independente de sua situação (evadido, diplomado, matriculado, trancado ou cancelado) e informações históricas dos discentes. Com a aplicação das tarefas de agrupamento, 6 grupos foram criados contendo elementos com maiores similaridade. Esta *view* permite uma análise de todos os discentes ingressados na instituição.

A Figura 14 apresenta um gráfico tridimensional representando os grupos formados com a aplicação da etapa de agrupamento.

A Figura 15 apresenta a contribuição dos atributos nos componentes, que representam os eixos PC1, PC2 e PC3 da Figura 14.

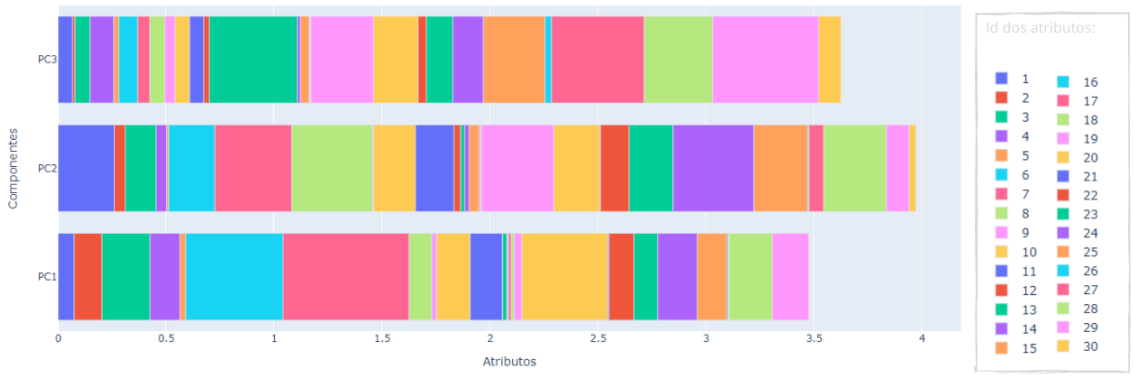
A Tabela 8 complementa o entendimento, apresentando a contribuição ($\alpha$) de cada atributo em cada componente. Os atributos que mais contribuem em cada componente estão enfatizados.

Figura 14 – Representação tridimensional do agrupamento dos dados da *view* 02




Fonte: Autor do trabalho

Figura 15 – Contribuição dos atributos da *view* 2 em cada componente




Fonte: Autor do trabalho

Tabela 8 – Detalhamento da contribuição de cada atributo nos componentes PC1, PC2 e PC3 da *view* 2

| **Id** | **Atributo** | $\alpha$ **PC1** | $\alpha$ **PC2** | $\alpha$ **PC3** |
|---|---|---|---|---|
| 1 | SEXO | 7% | 26% | 7% |
| 2 | COR DA PELE | 13% | 5% | 1% |
| 3 | TIPO ESCOLA | *22%* | 14% | 7% |
| 4 | TURNO DE INGRESSO | 14% | 5% | 11% |

| **Id** | **Atributo** | $\alpha$ **PC1** | $\alpha$ **PC2** | $\alpha$ **PC3** |
|---|---|---|---|---|
| 5 | MODE DE ADMISSAO | 3% | 1% | 2% |
| 6 | USO DE POLITICA AFIRMATIVA | **45%** | 21% | 9% |
| 7 | PARTICIPAÇÃO DE POLITICAS AFIRMATIVAS | **58%** | **36%** | 6% |
| 8 | MODALIDADE CONCORRENCIA | 11% | **37**% | 7% |
| 9 | MODALIDADE HOMOLOGADA | 2% | 0% | 5% |
| 10 | COD CURSO ATUAL | 16% | 19% | 7% |
| 11 | COD CURSO ADMISSÃO | 15% | 18% | 7% |
| 12 | TRANSFERENCIA INTERNA | 0% | 3% | 3% |
| 13 | SEMESTRE ADMISSAO | 2% | 2% | **41%** |
| 14 | IDADE INGRESSO | 0% | 2% | 1% |
| 15 | CIDADE ACADEMICO | 0% | 4% | 4% |
| 16 | PAIS FAMILIAR | 0% | 0% | 0% |
| 17 | ESTADO FAMILIAR | 1% | 1% | 0% |
| 18 | CIDADE FAMILIAR | 1% | 0% | 0% |
| 19 | CHECK ENDEREÇO | 4% | 34% | **29%** |
| 20 | TIPO VESTIBULAR | **40%** | 21% | 21% |
| 21 | DISTANCIA | 1% | 0% | 0% |
| 22 | PERCENTUAL CURSADO | 12% | 13% | 4% |
| 23 | TOTAL DISCIPLINA | 11% | 20% | 12% |
| 24 | PERCENTUAL APROVADO | 18% | **37**% | 14% |
| 25 | PERCENTUAL REPROVADO | 14% | 25% | **29**% |
| 26 | PERCENTUAL TRANCADO | 1% | 1% | 3% |
| 27 | PERCENTUAL CANCELADO | 0% | 7% | **43%** |
| 28 | PERCENTUAL EXAME FINAL | 20% | 29% | **32**% |
| 29 | PERCENTUAL DE DESISTENCIA NOS EXAMES FINAIS | 17% | 11% | **49%** |
| 30 | COEFICIENTE | 0% | 3% | 10% |

Fonte: Autor do trabalho

A seguir são elencadas as características descobertas, após as análises das tabelas e figuras anteriores, referentes a cada grupo.

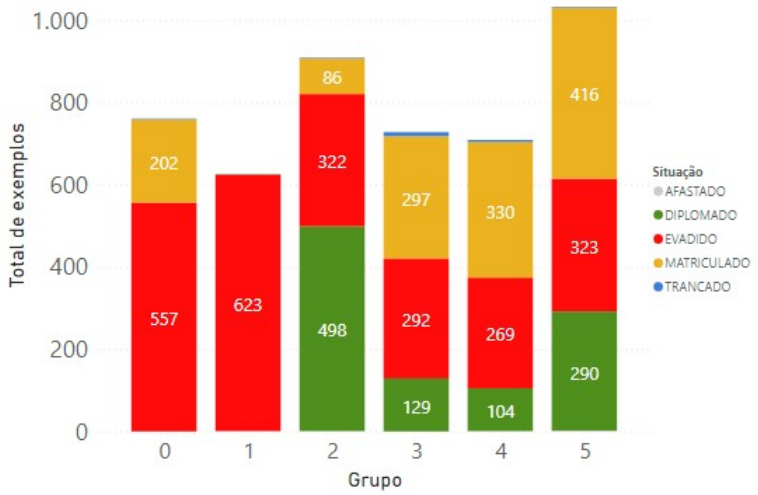
- **Grupo 0**: destaca-se com alta concentração de evadidos; dos 761 exemplos, 557 são discentes evadidos e 202 são discentes matriculados; esta configuração chama a atenção pois, esses 202 discentes matriculados foram agrupados juntos aos 557 evadidos; a similaridade entre estes discentes matriculados e discentes evadidos

faz entender que estes matriculados possuem características muito parecidas com discentes já evadidos, significando um ponto de atenção por parte da UFOP.

- **Grupo 1**: destacado com a cor vermelha na Figura 14, possui 627 discentes e deste total 623 são discentes evadidos; este grupo reúne discentes em sua maioria com o percentual cursado equivalente a 0%; provavelmente discentes com matrículas homologadas que evadiram ou discentes que evadiram no primeiro semestre.
- **Grupo 2**: destacado com a cor verde na Figura 14, reúne em maior parte discentes diplomados; observa-se que os evadidos contidos neste grupo possuem média de coeficiente e média do percentual de aprovação maior, em relação ao grupo 0 e grupo 1; estes 86 discentes possuem características similares a discentes diplomados.
- **Grupo 3**: contém discentes de várias situações; os discentes evadidos reunidos neste grupos possuem rendimentos significativamente bons; a média de coeficiente destes discentes é de 30,04 pontos.
- **Grupo 4**: reúne em sua grande maioria discentes matriculados; os discentes evadidos contidos no grupo são os que, mesmo com rendimentos bons, evadiram nos primeiros anos da graduação.
- **Grupo 5**: reúne discentes com maiores percentuais de aprovações e permanência; o grupo reúne 1033, e em sua maioria discentes matriculados; os discentes evadidos presentes neste agrupamento, encontram-se aqueles com melhores rendimentos.

Na Figura 16 pode-se observar a distribuição das situações dos discentes. Se um aluno matriculado, está inserido em um grupo com alta concentração de evadidos, significa que este aluno necessita-se de atenção especial, pois possui características e fatores similares aos de discentes evadidos. Sendo assim, alunos matriculados que pertencem ao grupo 0, precisam de atenção especial, assim como os grupos 3 e 4, que apesar de possuir alunos diplomados, o número de alunos evadidos superam o mesmo. A Tabela 9 apresenta algumas características de cada grupo.

Figura 16 – Distribuição das situações dos discentes entre os grupos gerados sobre a *view* 2




Fonte: Autor do trabalho

Tabela 9 – Análise dos grupos gerados a partir da *view* 2

| ***Parâmetro*** | **Grupo 0** | **Grupo 1** | **Grupo 2** | **Grupo 3** | **Grupo 4** | **Grupo 5** |
|---|---|---|---|---|---|---|
| Cor do grupo | Amarelo | Vermelho | Verde | Azul | Roxo | Rosa |
| Total de exemplos | 761 | 627 | 909 | 728 | 709 | 1033 |
| Média de coeficiente | 17,89 | 81 | 54,55 | 44,63 | 39,91 | 54,24 |
| Distância média entre a cidade da família à João Monlevade | 121 | 169 | 169 | 75 | 90 | 153 |
| Idade média de ingresso | 23 | 21 | 22 | 21 | 21 | 20 |
| Percentual cursado médio | 7% | 1% | 73% | 38% | 35% | 56% |
| Percentual médio de aprovação na disciplinas | 21% | 2% | 73% | 64% | 59% | 68 % |

Fonte: Autor do trabalho

A Tabela 10 contém o detalhamento das características dos discentes evadidos

reunidos em cada um dos grupos. A avaliação de fatores como média dos coeficientes, permanência e aprovação auxilia na distinção dos discentes evadidos contidos em um grupo para outro. O grupo 0 se difere do grupo 1 por ter alunos matriculados. Perceba que os grupos 1 e 2 se contrastam em relação ao percentual de aprovação, coeficiente e de tempo cursado. Os grupos 3, 4 são semelhantes, como visto na figura anterior. O grupo 5 é o mais diverso, com uma boa divisão entre todas as situações.

Tabela 10 – Detalhamento das características dos discentes evadidos contidos nos grupos da *view* 2

| **Parâmetro** | **Grupo 0** | **Grupo 1** | **Grupo 2** | **Grupo 3** | **Grupo 4** | **Grupo 5** |
|---|---|---|---|---|---|---|
| Cor do grupo | Amarelo | Vermelho | Verde | Azul | Roxo | Rosa |
| Total de exemplos | 557 | 623 | 322 | 292 | 269 | 323 |
| Coeficiente médio | 17,44 | 0,74 | 37,73 | 30,04 | 29,52 | 41,21 |
| Percentual cursado médio | 5,68% | 0,65% | 29,53% | 11,98% | 11,82% | 19,75% |
| Percentual de aprovação | 17,4% | 1,9% | 45,6% | 36,6% | 34,6% | 44,9% |

Fonte: Autor do trabalho

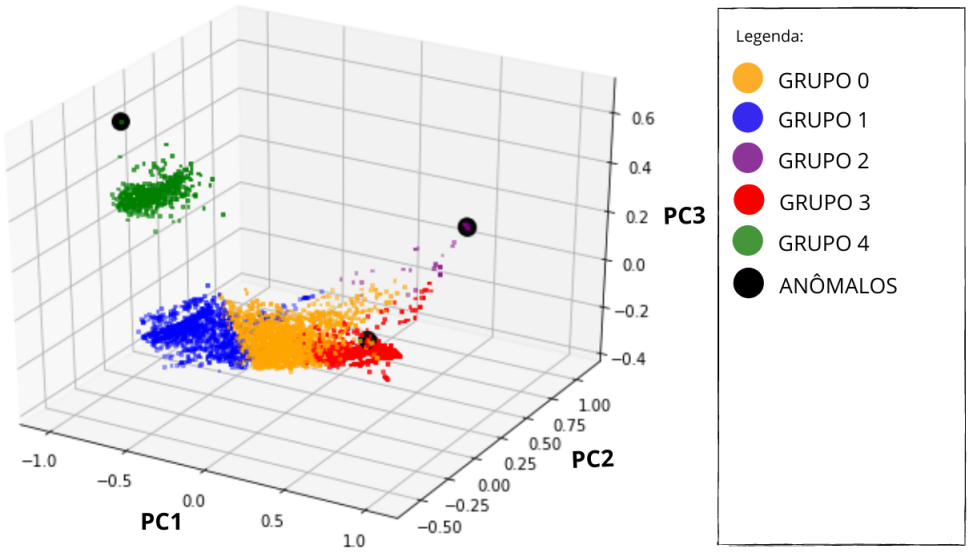
A análise desta *view* retorna um resultado satisfatório, levando em consideração os objetivos desta pesquisa. Com este conjunto, foi possível agrupar os discentes com características simulares em seis grupos. Os grupos reúnem discentes em várias situações, porém, percebe-se uma diferença quando é comparado discentes da mesma situação em grupos distintos. Quando são analisadas as características de cada grupo, é possível criar pontos de atenção voltados aos discentes matriculados e sua similaridade com o grupo que pertence. Se um aluno matriculado, está inserido em um grupo com alta concentração de evadidos, significa que este aluno necessita-se de atenção especial, pois possui características e fatores similares aos de discentes evadidos.

### 3.5.3 Apresentação dos resultados da *view* 03

Esta *view* contém parâmetros relacionados ao desempenho dos discentes nas disciplinas por eles realizadas. Com a técnica de agrupamento, discentes com rendimentos similares foram reunidos no mesmo grupo. Cinco grupos foram gerados e a Figura 17

contém a representação gráfica dos resultados do agrupamento.

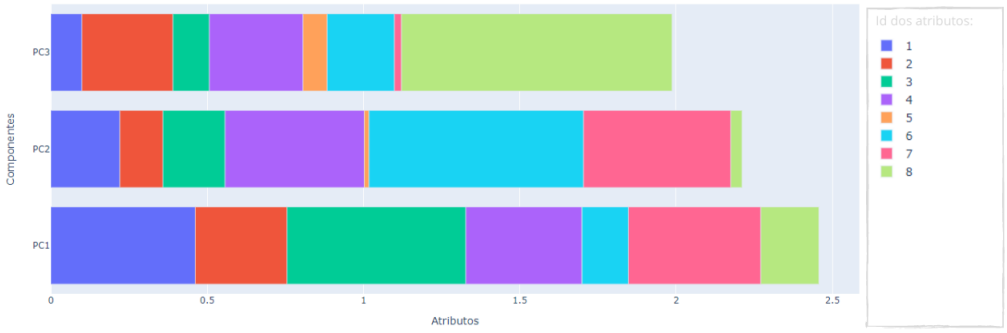
Figura 17 – Representação tridimensional do agrupamento dos dados da *view* 03




Fonte: Autor do trabalho

A Tabela 11 contém a contribuição de cada atributo desta *view* em cada componente gerado. A Figura 18 complementa o entendimento, apresentando a contribuição ($\alpha$) dos atributos em cada componente. Os atributos que mais contribuem em cada componente estão enfatizados em negrito.

Figura 18 – Contribuição dos atributos da *view* 3 em cada componente




Fonte: Autor do trabalho

Tabela 11 – Detalhamento da contribuição de cada atributo nos componentes PC1, PC2 e PC3 da *view* 3

| **Id** | **Atributo** | $\alpha$ **PC1** | $\alpha$ **PC2** | $\alpha$ **PC3** |
|---|---|---|---|---|
| 1 | COEFICIENTE | **46%** | 22% | 10% |
| 2 | TOTAL DISCIPLINA | 29% | 14% | **29%** |
| 3 | PERCENTUAL APROVADO | **57%** | 20% | 12% |
| 4 | PERCENTUAL REPROVADO | 37% | **45%** | **30**% |
| 5 | PERCENTUAL TRANCADO | 0% | 2% | 8% |
| 6 | PERCENTUAL CANCELADO | 15% | **69%** | 22% |
| 7 | PERCENTUAL DE USO EM EXAME FINAL | **42%** | **47%** | 2% |
| 8 | PERCENTUAL CURSADO | 19% | 4% | **87%** |

Fonte: Autor do trabalho

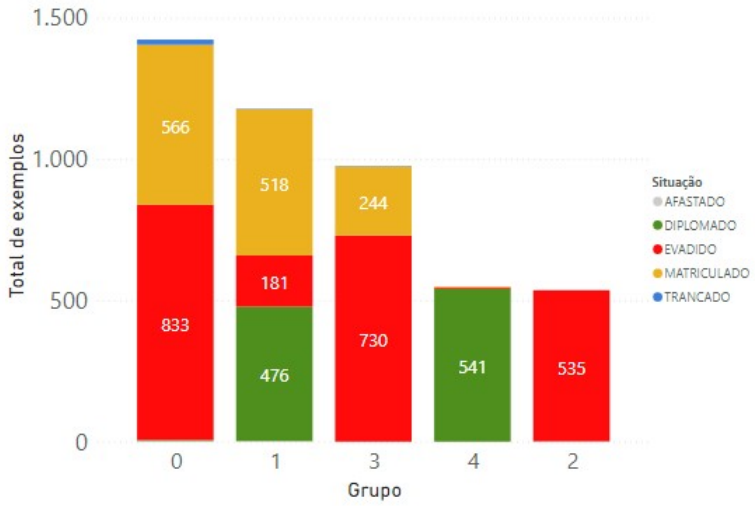
A seguir são elencadas as características descobertas, após as análises das tabelas e figuras anteriores, referentes a cada grupo.

- **Grupo 0**: possui uma 833 discentes evadidos e 566 discentes matriculados; o fato de um aluno matriculado ser agrupado em um grupo com discentes evadidos de baixos rendimentos significa que estes possuem uma similaridade com discentes que vieram a evadir.
- **Grupo 1**: já possui uma concentração de discentes diplomados e matriculados e os 181 discentes evadidos deste grupo possuem melhores rendimentos e maior permanência na instituição, vindo a evadir possivelmente por desligamentos por prazo máximo ou motivos particulares; os discentes matriculados contidos neste grupo possuem rendimentos considerado satisfatórios.
- **Grupo 2**: reúne apenas discentes evadidos, com exceção de 1 aluno afastado; estes discentes evadidos provavelmente desistiram nos primeiros semestres.

- **Grupo 3**: contem 730 discentes evadidos; em sua maioria são discentes que saíram da graduação nos primeiros semestres, como se percebe ao olhar o percentual médio cursado, equivalente a 5%; destaca-se que neste grupo 244 discentes matriculados que possuem similaridade com discentes que vieram a evadir.
- **Grupo 4**: reúne em sua grande maioria discentes diplomados, com raras exceções; o percentual de aprovação nas disciplinas equivale a 84% e uma média de coeficiente de 70,01 pontos.

Com esta análise realizada sobre a *view* 3, foi possível agrupar os discentes com rendimentos similares. Analisando apenas os discentes evadidos, foi possível perceber a diferença entre os discentes e os grupos. Determinados grupos possuem discentes, mesmo evadidos, com rendimentos considerados satisfatórios, similares aos de discentes diplomados. Outros concentram discentes evadidos com rendimentos não satisfatórios. Com a aplicação desta técnica, foi possível perceber entre os discentes matriculados, aqueles que possuem similaridades com discentes que vieram a evadir, além daqueles que possuem similaridade com discentes diplomados.

Figura 19 – Distribuição das situação dos discentes entre os grupos gerados sobre a *view* 03




Fonte: Autor do trabalho

A Tabela 12 apresenta algumas características dos grupos formados.

Tabela 12 – Análise dos grupos gerados a partir da *view* 3

| ***Parâmetro*** | **Grupo 0** | **Grupo 1** | **Grupo 2** | **Grupo 3** | **Grupo 4** |
|---|---|---|---|---|---|
| Cor representativa | Amarela | Azul | Roxo | Vermelho | Verde |

| ***Parâmetro*** | **Grupo 0** | **Grupo 1** | **Grupo 2** | **Grupo 3** | **Grupo 4** |
|---|---|---|---|---|---|
| Total de exemplos | 1423 | 1179 | 536 | 976 | 549 |
| Média de coeficiente | 37,12 | 64,45 | 0,52 | 14,73 | 70,01 |
| Percentual de aprovação nas disciplinas | 41% | 78% | 0,8% | 13% | 84% |
| Percentual cursado médio | 22% | 75% | 1% | 5% | 104% |
| discentes evadidos | 833 | 181 | 535 | 730 | 5 |
| discentes diplomados | 4 | 476 | 0 | 0 | 541 |
| discentes matriculados | 566 | 518 | 0 | 244 | 3 |

Fonte: Autor do trabalho

# 4 Conclusão

Através das técnicas de mineração de dados foi possível caracterizar a evasão de discentes, reunindo discentes com características similares. Esta caracterização, sem o auxílio de técnicas computacionais, seria complexa de ser realizada.

Entre os principais achados, destacam-se a criação de cinco grupos de discentes evadidos. Cada grupo formado possui características distintas. Há grupos que reúnem discentes que evadem no primeiro período da graduação, sem se quer concluir as disciplinas do primeiro semestre. Outros reúnem discentes que permaneceram alguns semestres na universidade e vieram a evadir, percebendo um rendimento não satisfatório nestes grupos. Existe também um grupo formado onde, discentes possuem rendimentos considerados satisfatórios, similares a diplomados, mas por motivos como o desligamento por prazo máximo evadem. Ao observar as características gerais dos discentes reunidos em cada grupo, pode-se perceber diferenças entre os discentes evadidos. Com a criação destes grupos, é possível rotular discentes evadidos em relação à evasão.

Outro achado foi obtido ao aplicar a tarefa de agrupamento em uma base com todos os discentes do instituto, independente de sua situação. Após o agrupamento, é possível comparar em qual grupo os discentes matriculados estão inseridos. Percebe-se que existem discentes matriculados que possuem similaridade com discentes evadidos, ao qual foram reunidos no mesmo grupo. O fato de um discente matriculado possuir similaridade com discentes evadidos é um ponto de atenção, ao qual os responsáveis da instituição podem estudar estrategias para impedir que estes matriculados venham a evadir. Nesta analise também, percebe-se que cada grupo reúne evadidos com características distintas entre os demais.

A detecção dos discentes anômalos possibilita encontrar os casos de alunos evadidos que possuem maior anormalidade. Entre os casos anômalos estão discentes com rendimentos satisfatórios, mas evadiram por prazo máximo. Estes discentes possuem rendimentos, percentuais de aprovação e permanência de grupo similares a alunos evadidos, mas atingiram o limite para se graduarem.

Outro ponto importante está na possibilidade de predizer em qual grupo um novo aluno pertencerá, conforme o módulo *predict* descrito no Apêndice A. Com esta tarefa, é possível analisar em qual grupo este novo exemplo será inserido. Assim, pode-se obter algumas conclusões, observando as características deste grupo que ele se enquadrou.

Destaca-se que uso das técnicas de mineração de dados demanda mão de obra com um nível de conhecimento específico voltado ao tema para o desenvolvimento e aplicação. Levando em consideração esse ponto, o ICEA se destaca por se tratar de um

dos institutos referência em ensino e pesquisa voltadas às áreas tecnológicas, contendo alunos e professores com conhecimento aprofundado voltados à área. Este fator, além de proporcionar ganhos aos alunos pela experiência no desenvolvimento de pesquisas, iniciação científica ou extensões, retorna ao instituto resultados vantajosos com tecnologia de última geração. Esta pesquisa fomenta a continuidade deste trabalho voltado à problemática da evasão, além da abordagens de outros problemas.

## 4.1 Trabalhos futuros

A seguir, são apresentados possíveis pontos de melhorias, sugestões trabalhos futuros de continuidade baseados nesta pesquisa. A continuidade do trabalho, poderá aumentar ainda mais a obtenção de conhecimento em relação à evasão, além de outros temas. Mesmo atendendo os requisitos propostos, há pontos de melhorias que podemos ser desenvolvidos em trabalhos futuros.

O desenvolvimento de formas automatizadas e otimizadas para atualização das fontes existentes com dados dos anos seguintes é uma delas. Melhorar este processo possibilitará o prosseguimento das análises realizadas nesta pesquisa de forma mais rápida e simplificada. Aponta-se também, a inserção de novos dados, vindo de fontes internas ou externas, poderão agregar e otimizar ainda mais os resultados das análises desenvolvidas. A seguir, existem algumas sugestões de novas fontes:

- Respostas dos formulários oficiais da UFOP: a UFOP disponibiliza formulários analisando as satisfação dos alunos nos finais dos semestres letivos. As respostas destes formulários é uma das possíveis fontes, analisando a satisfação dos alunos em relação curso, professores, estruturas, e vincular estes dados com a situação do aluno poderá ser uma fonte com potencial para gerar conhecimento útil.
- Avaliação socioeconômica dos alunos: o Núcleo de Assuntos Comunitários Estudantis (NACE) realiza avaliações socioeconômicas dos alunos, contendo uma análise criteriosa. As fontes de dados deste núcleo contém uma gama de informações sensíveis, demandando critérios de segurança dos dados.
- Dados do Exame Nacional de Desempenho dos Estudantes (ENADE): os resultados do ENADE é uma possível fonte de dados, principalmente para análise dos alunos diplomados.

Este trabalho pode ser fonte de embasamento para vários outras abordagens. Esta pesquisa foi voltada à problemática da evasão, mas trabalhos futuros poderão ser desenvolvidos abordando os alunos diplomados, matriculados ou retidos na universidade, além da aplicação da análise de evasão em outros institutos da UFOP.

A aplicação de outras tarefas de mineração de dados, como a tarefa de classificação, regressão e determinação de regras de associação para obtenção de conhecimento também são apontadas como possíveis sugestões de trabalhos futuros.

# Referências


Bardagi, M. P.: 2007, Evasão e comportamento vocacional de universitários: estudo sobre desenvolvimento de carreira na graduação. Citado na página 16.

Belenke dos Santos, J. C.: 2021, Usando mineração de dados para predição da evasão escolar. Citado na página 16.

Bispo, C. A. F.: 1998, Uma análise da nova geração de sistemas de apoio à decisão, *São Carlos* . Citado na página 22.

Braga, L. P. V. B.: 2005, *Introdução à Mineração de Dados-2a edição: Edição ampliada e revisada*, Editora E-papers. Citado na página 26.

Calil, L. A. d. A., Carvalho, D. R., Santos, C. B. d. and Vaz, M. S. M. G.: 2008, Mineração de dados e pós-processamento em padrões descobertos. Citado na página 25.

Camilo, C. O. and Silva, J. C. d.: 2009, Mineração de dados: Conceitos, tarefas, métodos e ferramentas, *Universidade Federal de Goiás (UFC)* pp. 1–29. Citado 3 vezes nas páginas 25, 26 e 27.

Castanheira, L. G.: 2008, Aplicação de técnicas de mineração de dados em problemas de classificação de padrões, *Belo Horizonte: UFMG* . Citado na página 16.

Castro, A. and Gouvêa, G.: 2014, Identificação dos fatores que influenciam no tempo até a evasão/retenção dos alunos do iceb-ufop utilizando técnicas paramétricas em análise de sobrevivência (pp. 820-824), *Revista da Estatística da Universidade Federal de Ouro Preto* 3(3). Citado na página 25.

Coelho, F. C.: 2007, *Computação Científica com Python*, Lulu. com. Citado na página 28.

De Amo, S.: 2004, Técnicas de mineração de dados, *Jornada de Atualizaçao em Informatica* . Citado 2 vezes nas páginas 26 e 27.

DIAS, E. C., THEÓPHILO, C. R. and 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, São Paulo, SP*, Vol. 7. Citado 3 vezes nas páginas 16, 22 e 24.

Do Carmo, A. J.: 2018, *EVASÃO UNIVERSITÁRIA: REPERCUSSÕES NA TRAJETÓRIA E NO PROJETO DE VIDA DOS JOVENS*, PhD thesis, Universidade Federal de Viçosa. Citado na página 24.

dos Santos, A. P.: 1999, Diagnóstico do fluxo de estudantes nos cursos de graduação da ufop: retenção, diplomação e evasão, *Avaliação-Revista da Avaliação da Educação Superior* 4(4). Citado na página 24.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P.: 1996, From data mining to knowledge discovery in databases, *AI Magazine* 17(3), 37.
**URL:** *https://ojs.aaai.org/index.php/aimagazine/article/view/1230* Citado na página 26.

Fernandes, K. C.: 2019, Estudo da evasão de alunos de graduação utilizando educational data mining. Citado na página 28.

Gonçalves, T. C., da Silva, J. C. and Cortes, O. A. C.: 2018, Técnicas de mineração de dados: um estudo de caso da evasão no ensino superior do instituto federal do maranhão, *Revista Brasileira de Computação Aplicada* 10(3), 11–20. Citado 2 vezes nas páginas 22 e 25.

Junior, P. L., Bisinoto, C., de Melo, N. S. and Rabelo, M.: 2019, Taxas longitudinais de retenção e evasão: uma metodologia para estudo da trajetória dos estudantes na educação superior, *Ensaio: Avaliação e Políticas Públicas em Educação* 27(102), 157–178. Citado na página 17.

Machado, R. D., Nara, E. O. B., Schreiber, J. N. C. and Schwingel, G. A.: 2015, Estudo bibliométrico em mineração de dados e evasão escolar, *Apresentado na XI Congresso Nacional de Excelência em Gestão, Rio de Janeiro, RJ*. Citado na página 16.

Passos, E. O.: 2016, Evasão e diplomação no curso de administração da ufop. Citado na página 25.

Silva Filho, R. L. L., Motejunas, P. R., Hipólito, O. and Lobo, M. B. C. M.: 2007, A evasão no ensino superior brasileiro., *Cadernos de pesquisa* 37(132), 641–659. Citado na página 22.

Varella, C. A. A.: 2008, Análise de componentes principais, *Seropédica: Universidade Federal Rural do Rio de Janeiro* . Citado na página 38.

Veloso, T. C. M. and de Almeida, E. P.: 2013, Evasão nos cursos de graduação da universidade federal de mato grosso, campus universitário de cuiabá–um processo de exclusão, *Série-Estudos-Periódico do Programa de Pós-Graduação em Educação da UCDB* . Citado na página 16.

Witten, I. H., Frank, E., Hall, M. A. and Pal, C. J.: 2017, *Data Mining Practical Machine Learning Tools and Techniques*, fourth edi edn, Morgan Kaufmann. Citado na página 27.

# Anexos

# ANEXO A – Códigos de implementação

## A.1 Processamento das *views*

Os arquivos gerados em formatos de planilhas foram salvos em uma pasta no serviço de armazenamento e sincronização de arquivos do *google drive*, em seguida conectados ao *google Colaboratory*.

O código abaixo exemplifica a etapa de conexão e preparo dos *views* salvos no *google drive*. Contam-se com dois *dataframes*, um recebe os dados e outro recebe apenas o identificador criado e utilizado na exportação dos resultados do processamento dos algoritmos.

```
# BIBLIOTECA UTILIZADA:
import pandas as pd
# CONEXAO DO DATASET
url = "/content/drive/MyDrive/TCC II/DATASETS/DATASET_05_NOTAS_ALUNOS.xlsx"
dataFrame = pd.read_excel(url)
#PREPARO DOS RESULTADOS
resultado = pd.DataFrame()
resultado['ID'] = dataFrame['ID']
dataFrame = dataFrame.drop(['ID'], axis=1)
```

## A.2 Transformação dos dados simbólicos e normalização

O código a seguir apresenta a aplicação das etapas de transformação dos dados simbólicos e normalização no *dateset* referente aos dados gerais dos alunos. Inicia-se com as importações das bibliotecas necessárias, seguindo com a montagem do *dateset*. Uma lista é criada, recebendo o nome de todos os atributos do tipo texto para conversão. A etapa de transformação dos dados simbólicos é feita através de um laço de repetições, realizando o processo em todos os valores necessários da *view*. Ao final da execução deste laço, todos os valores do conjunto de dados são de tipos numéricos. Por fim, é realizado a normalização.

```
# Bibliotecas utilizadas:
import pandas as pd
from sklearn import preprocessing

# Dataset:
url = "/content/BASE_DADOS_ALUNOS.xlsx"
```

```
dataFrame = pd.read_excel(url)

# Transformacao dos dados simbolicos em dados numericos:
lista_colunas_tipo_simbolico = ['DESCRICAO_MODO_ADMISSAO', 'TURNO', '
    TIPO_ESCOLA']
le = preprocessing.LabelEncoder()
for x in lista_colunas_tipo_simbolico:
    le.fit(dataFrame[x])
    dataFrame[x] = le.transform(dataFrame[x].astype(str))

# Normalizacao dos dados:
df = dataFrame
dataFrame = (df-df.min())/(df.max()-df.min())
```

## A.3 Determinação dos melhores parâmetros

O código abaixo, apresenta a análise do melhor número de grupos através do método de *Elbow*:

```
# BIBLIOTECAS NECESSRIAS
from sklearn.cluster import KMeans

# MTODO ELBOW
var = []

for i in range(1, 10):
    kmeans = KMeans(n_clusters = i, init = 'random')
    kmeans.fit(dataFrame)
    print (i,kmeans.inertia_)
    var.append(kmeans.inertia_)

# PLOTAGEM DOS RESULTADOS
plt.plot(range(1, 10), var)
plt.xlabel('Numero de clusters')
plt.ylabel('Varianca dos dados')
plt.show()
```

Descrevendo o código de forma sucinta, um vetor vazio é responsável por receber os resultados da variança dos dados. Em um laço de repetições, o valor da variança é calculado para k valor, iniciando com k = 1. A plotagem dos resultados é expressa em um gráfico bidimensional, onde o eixo x contem o número de *clusters* e o eixo y contém os valores da variança.

## A.4 Algoritmo *KMeans* e módulo *predict*

O código apresentado a seguir contém o detalhamento da aplicação do algoritmo *KMeans*. Inicialmente os parâmetros são ajustados, permitindo configurações. Por meio do método *fit*, o conjunto de dados é agrupado. Os grupos são inseridos em um *dataframe* destinado à exportação dos resultados.

```
# BIBLIOTECAS NECESSARIAS
from sklearn.cluster import KMeans

# AGRUPAMENTO
algorithm = KMeans(
    algorithm='auto',
    copy_x=True,
    init='k-means++',
    max_iter=3000,
    n_clusters=5,
    n_init=5,
    n_jobs=None,
    precompute_distances='auto',
    random_state=None,
    tol=0.0001,
    verbose=0)

algorithm.fit(dataFrame)

# MONTAGEM DOS RESULTADOS
resultado['grupo'] = algorithm.labels_
```

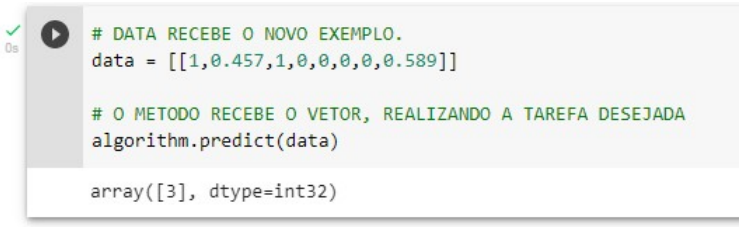
Pode-se predizer o grupo de um novo exemplo, através do método *predidict* da biblioteta *scikit-learn*. O método recebe um vetor, contendo um ou vários exemplos e retorna como resultado, qual grupo este novo elemento será incluído. Ressaltando apenas que para inserção de novos exemplos, os dados precisam estar normalizados.

```
# DATA RECEBE O NOVO EXEMPLO.
data = [[1,0.457,1,0,0,0,0,0.589]]

# O METODO RECEBE O VETOR, REALIZANDO A TAREFA DESEJADA
algorithm.predict(data)
```

Com esta tarefa, é possível inserir novos exemplos e analisar as características do grupo que ele foi inserido. A Figura 20 apresenta um exemplo da execução desta tarefa. Um exemplo foi inserido para predição, retornando que ele pertencerá ao grupo 3.

Figura 20 – Demonstração da execução da tarefa de predição




Fonte: Autor do trabalho

## A.5 Algoritmo *NearestNeighbors*

A seguir temos a apresentação do código relacionado à aplicação do algoritmo *NearestNeighbors*:

```
# BIBLIOTECAS NECESSARIAS
from sklearn.neighbors import NearestNeighbors

# DETECCAO DE ANOMALIAS
# MONTAGEM DO MEDOLO
nbrs = NearestNeighbors(n_neighbors = 50)
# fit model
nbrs.fit(dataFrame)

distances, indexes = nbrs.kneighbors(dataFrame)
# REPRESENTACAO GRAFICA DAS MEDIAS DAS DISTISTANCIAS DE CADA EXEMPLO
fig = plt.figure(figsize = (9,7))
plt.plot(distances.mean(axis =1))

# FILTROS DOS VALORES ANOMALOS
corte = 0.5
plt.axhline(y=0.2, xmin=0.04, xmax=0.96, color='red')
outlier_index = np.where(distances.mean(axis = 1) > corte)
outlier_values = dataFrame.iloc[outlier_index]

# PREPARO DOS RESULTADOS
anomalos = resultado.iloc[outlier_index]
```

Inicialmente o modelo é ajustado, passando o número de vizinhos. A plotagem gráfica auxilia na escolha do número de viznhos k o valor atributos à variável corte, responsável por filtrar os exemplos discrepantes. Para melhor escolha do número de vizinhos, referentes ao valor k e o valor da variável corte, foram realizados vários testes.

O *dataframe* nomeado como "anomalos', recebe o índice de identificação dos exemplos considerado anômalos.

## A.6 Técnica PCA

A seguir, temos a apresentação do código que realiza esta tarefa. Para o uso da técnica de PCA nesta pesquisa, foi utilizado o método *decomposition* da biblioteca do *sklearn*. O número de componentes é passado como parâmetro.

```
# Bibliotecas utilizadas:
from sklearn.decomposition import PCA

pca = PCA(n_components=3)
pca_alunos = pca.fit_transform(dataFrame_normalizado)
pca_alunos_df = pd.DataFrame(data = pca_alunos, columns = ["Componente_1",
    "Componente_2", "Componente_3"])
pca_nome_alunos = pd.concat([pca_alunos_df, resultado['grupo']], axis =1)
```