INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE MINAS GERAIS - *CAMPUS* SÃO JOÃO EVANGELISTA
BACHARELADO EM SISTEMAS DE INFORMAÇÃO

Elis Daimara Bresciani; Marcelo Pereira Gonçalves

USO DA MINERAÇÃO DE DADOS PARA IDENTIFICAR PADRÕES NO DESEMPENHO DOS ALUNOS DOS CURSOS SUPERIORES DO IFMG-SJE

SÃO JOÃO EVANGELISTA
2023

ELIS DAIMARA BRESCIANI
MARCELO PEREIRA GONÇALVES

USO DA MINERAÇÃO DE DADOS PARA IDENTIFICAR PADRÕES NO DESEMPENHO DOS ALUNOS DOS CURSOS SUPERIORES DO IFMG-SJE

Trabalho de conclusão de curso apresentado ao Curso Bacharelado em Sistemas de Informação do Instituto Federal de Minas Gerais - *Campus* São João Evangelista como exigência parcial para obtenção do título de Bacharel em Sistemas de Informação.

Orientador: Prof. Dr. Fábio Rodrigues Martins
Coorientador: Prof. Me. Rosinei Soares de Figueiredo

SÃO JOÃO EVANGELISTA
2023

B842u Bresciani, Elis Daimara.
Uso da mineração de dados para identificar padrões no desempenho dos alunos dos cursos superiores do IFMG - SJE [manuscrito] / Elis Daimara Bresciani, Marcelo Pereira Gonçalves. – 2023.
92 f. : il.

Orientador: Fábio Rodrigues Martins.
Coorientador: Rosinei Soares de Figueiredo.
Trabalho de Conclusão de Curso (bacharelado) – Instituto Federal de Minas Gerais. *Campus* São João Evangelista, 2023.

1. Mineração de dados (computação). 2. Indicadores de desempenho. 3. Educação - Tecnologia. I. Gonçalves, Marcelo Pereira. II. Martins, Fábio Rodrigues. III. Figueiredo, Rosinei Soares de. IV. Instituto Federal de Minas Gerais. *Campus* São João Evangelista. V. Título.

CDD: 371.3078

Catalogação: Kelly Cristiane Santos Morais - CRB-6/3217

ELIS DAIMARA BRESCIANI; MARCELO PEREIRA GONÇALVES

USO DA MINERAÇÃO DE DADOS PARA IDENTIFICAR PADRÕES NO DESEMPENHO DOS ALUNOS DOS CURSOS SUPERIORES DO IFMG-SJE

Trabalho de conclusão de curso apresentado ao curso bacharelado em Sistemas de Informação do Instituto Federal de Minas Gerais - *Campus* São João Evangelista como exigência parcial para obtenção do título de Bacharel em Sistemas de Informação.

Aprovado em 13/06/2023

BANCA EXAMINADORA

gov.br Documento assinado digitalmente
FABIO RODRIGUES MARTINS
Data: 28/07/2023 13:52:08-0300
Verifique em https://validar.iti.gov.br

Orientador Prof. Dr. Fábio Rodrigues Martins
Instituto Federal de Minas Gerais – Campus São João Evangelista

gov.br Documento assinado digitalmente
ROSINEI SOARES DE FIGUEIREDO
Data: 31/07/2023 09:29:37-0300
Verifique em https://validar.iti.gov.br

Coorientador Prof. Me. Rosinei Soares de Figueiredo
Instituto Federal de Minas Gerais – Campus São João Evangelista

gov.br Documento assinado digitalmente
EDUARDO AUGUSTO COSTA TRINDADE
Data: 31/07/2023 10:04:09-0300
Verifique em https://validar.iti.gov.br

Convidado: Prof. Me. Eduardo Augusto Costa Trindade
Instituto Federal de Minas Gerais – Campus São João Evangelista

Dedicamos essa monografia às nossas famílias, amigos e professores que nos apoiaram e nos incentivaram nessa jornada até aqui.

AGRADECIMENTOS

Primeiramente agradecemos a Deus, por nos acompanhar em todos os momentos e nos permitir mais esta conquista. Aos nossos pais, que com tanto amor, sabedoria e paciência nos criaram e educaram, bem como acreditaram e investiram em nossa vida profissional e aos nossos familiares, pelo carinho, encorajamento e acolhimento.

Ao “Clube da Luta”, pela alegria de caminhar com vocês, pelas risadas e conversas, pela torcida e paciência de nos ouvirem falar sobre este trabalho a cada passo, estando ao nosso lado nos momentos bons e ruins. Por fim, a todos os nossos amigos, por fazerem parte das nossas vidas e tornarem tudo muito mais feliz, o caminho até aqui foi (muito) longo e não teríamos conseguido sem vocês.

Agradecemos em especial aos nossos orientadores, pela oportunidade e pela ajuda neste trabalho. Agradecemos também aos professores e demais servidores do IFMG-SJE, que nos guiaram no decorrer desses anos e compartilharam conosco o conhecimento necessário para chegarmos até aqui, o que aprendemos nessa instituição é de um valor inestimável.

Enfim, agradecemos a todos que, de uma forma ou de outra, contribuíram no decorrer do curso e para a realização deste trabalho.

“E levante, gay! Que a luta ainda não acabou pras gay. Que a nossa vitória vai ser o close, gay! E que se eu tô aqui hoje dando voz pras gay, é por ser gay!”

(Gloria Groove)

RESUMO

O desempenho acadêmico avalia se os objetivos de aprendizagem foram atingidos, registrando a trajetória dos discentes no currículo escolar e desta forma possibilita verificar o nível de absorção de conhecimento pelos alunos. De acordo com o Ministério da Educação o bom desempenho é um importante mediador de aprendizagem e se relaciona com o posterior sucesso profissional, podendo ser impactado por diversos fatores internos e externos à instituição ao longo da vida acadêmica do aluno. Com base nisso, esse trabalho consiste em identificar padrões relacionados ao desempenho dos alunos, envolvendo os processos do KDD e técnicas de mineração de dados nas bases onde encontram-se os registros acadêmicos dos alunos dos cursos superiores do Instituto Federal de Minas Gerais – *Campus* São João Evangelista. Os itens que norteiam o desenvolvimento deste trabalho são os indicadores de desempenho educacionais de fluxo, pesquisas sobre KDD, assim como a mineração de dados educacionais. A metodologia para a realização desta pesquisa tem caráter qualitativo-quantitativo de natureza aplicada, tendo como o método de pesquisa dedutivo e aplicação da técnica de mineração de dados através do processo de KDD. Os procedimentos utilizados para executar a pesquisa são: observação, pesquisa bibliográfica, pesquisa ação e experimento. Foi possível identificar ao longo do desenvolvimento das análises alguns fatores relevantes e que contribuem para a desistência, permanência e conclusão dos alunos, como por exemplo, a quantidade de faltas e reprovações em matérias ao longo do curso, que se mostraram como fatores cruciais para a evasão.

Palavras-chave: Mineração de Dados; Indicadores de Desempenho; *Knowledge Discovery in Database.*

ABSTRACT

Academic performance evaluates whether the learning objectives were achieved, recording the trajectory of students in the school curriculum and thus makes it possible to verify the level of absorption of knowledge by students. According to the Ministry of Education, good performance is an important learning mediator and is related to later professional success, and can be impacted by various internal and external factors to the institution throughout the student's academic life. Based on this, this work consists of identifying patterns related to student performance, involving the KDD processes and data mining techniques in the databases where the academic records of students from higher education courses at the Federal Institute of Minas Gerais - Campus São Paulo are located. John Evangelist. The items that guide the development of this work are the educational performance indicators of flow and quality, research on KDD and data mining, as well as educational data mining. The methodology for carrying out this research has a qualitative-quantitative nature of applied nature, having as the hypothetical-deductive research method and application of the data mining technique through the KDD process. The procedures used to carry out the research are: observation, bibliographical research, action research and experiment. It was possible to identify throughout the development of the analyses some relevant factors that contribute to the dropout, permanence and completion of the students, such as the number of absences and failures in subjects throughout the course, which proved to be crucial factors for dropping out.

Keywords: Data Mining; Performance indicators; Knowledge Discovery in Database.

LISTA DE ILUSTRAÇÕES

LISTA DE ABREVIATURAS E SIGLAS

ABMES - Associação Brasileira de Mantenedoras de Ensino Superior

BD - Banco de Dados

BFS - *Breadth-first search*

CAPES - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior

CONAES - Comissão Nacional de Avaliação da Educação Superior

CPC - Conceito Preliminar de Curso

DCBD - Descoberta de Conhecimento em Base de Dados

EaD - Ensino a Distância

EDM - *Educational Data Mining*

Enade - Exame Nacional de Desempenho dos Estudantes

ENEM - Exame Nacional do Ensino Médio

GLP - *General Public License*

IDD - Desempenhos Observado e Esperado

IES - Instituição de Ensino Superior

IFMG-SJE - Instituto Federal de Minas Gerais campus São João Evangelista

IGC - Índice Geral de Cursos

INEP - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira

KDD - *Knowledge Discovery in Database*

LDB - Lei de Diretrizes e Bases da Educação Nacional

LGPD - Lei Geral de Proteção de Dados

MD - Mineração de Dados

PI's - *Performance Indicators*

PSF - *Python Software Foundation*

SC - Sem Conceito

SISU - Sistema de Seleção Unificada

SEMESP - Secretaria de Modalidades Especializadas de Educação

TICs- Tecnologias da Informação e Comunicação

UF - Unidade Federativa

UFF - Universidade Federal Fluminense

UFPR - Universidade Federal do Paraná

UNISC - Universidade de Santa Cruz do Sul

SUMÁRIO

1. INTRODUÇÃO

O Brasil possui 8,6 milhões de estudantes matriculados na graduação (cerca de 4% dos seus habitantes), distribuídos em 2.457 instituições de educação superior instaladas em todas as unidades da Federação. Em 2020, o número de concluintes em cursos de graduação presencial teve queda de 6,0% em relação a 2019. A modalidade a distância teve um aumento significativo de 26,7% no mesmo período (BRASIL, 2020).

Segundo Brasil (2020), ao considerar como fator de pesquisa o grau acadêmico, percebeu-se que o número de concluintes nos cursos tecnológicos foi o único a ter um aumento, cerca de 20,8% em 2020, quando comparado a 2019. Os graus de Bacharelado e Licenciatura registraram quedas de -0,9% e -4,2%, respectivamente no mesmo período.

A Secretaria de Modalidades Especializadas de Educação (SEMESP), representante das mantenedoras de ensino superior no Brasil, divulgou dados de 2021, em que a taxa de evasão chegou aos 36,6%, equivalente a 3,42 milhões de alunos considerando as modalidades de ensino à distância e presencial. O resultado foi ainda pior em 2020, quando 3,78 milhões de alunos largaram seus cursos.

O Censo da Educação Superior coleta os dados das instituições de ensino superior (IES) para fazer um acompanhamento e análise através dos indicadores de fluxo, desde quando o aluno entra no curso de graduação, até a sua saída, seja ela por meio da formatura, ou a desistência.

> Os indicadores de fluxo servem de base para diversas análises e também para medida da eficiência de cada curso, podendo auxiliar na criação de novos parâmetros de controle de eficiência do curso e na capacidade deste em formar pessoas, além de qualificar a oferta e a demanda desses cursos. Os indicadores de fluxo são: a permanência, a desistência e a conclusão no curso de ingresso (MINISTÉRIO DA EDUCAÇÃO, 2020a).

Conforme o Ministério da Educação (2020a), além dos métodos relacionados ao fluxo, existem cinco indicadores adicionais que são importantes instrumentos de avaliação da educação superior brasileira. Esses indicadores são: o Conceito do Exame Nacional de Desempenho dos Estudantes (Enade), o Indicador de Diferença entre os Desempenhos

Observado e Esperado (IDD), o Conceito Preliminar de Curso (CPC), o Índice Geral de Cursos (IGC) e o Indicador de Resultados. Esse conjunto de indicadores formam os Indicadores de Qualidade da Educação Superior, estando relacionados diretamente com o Ciclo Avaliativo do Enade.

Através dos indicadores de desempenho, sejam de fluxo ou qualidade apresentados pelo Ministério da Educação (2020a), tem-se uma dimensão clara da qualidade de uma Instituição de Ensino Superior (IES). No entanto, esses indicadores, mensuram de forma abrangente e em conjunto o desempenho dos alunos, sem entrar em um contexto mais detalhado de fatores específicos que possam impactá-los. Neste sentido, é importante desenvolver métodos para acompanhar o desempenho dos alunos.

Castro e Ferrari (2016) argumentam que através das técnicas de mineração de dados é possível realizar análises que ajudam na identificação de padrões relevantes. Dessa forma, tem-se o conhecimento e informações necessárias a partir de grandes volumes de dados, que servem principalmente para a tomada de decisão, nesse caso, referente ao desempenho dos alunos e das instituições de ensino. Os resultados podem ser analisados através de gráficos e tabelas, onde os conjuntos de dados analisados e os padrões identificados estarão dispostos de forma visual.

> A utilização da mineração de dados atraiu muita atenção na indústria da informação e na sociedade como um todo nos últimos anos, devido à ampla disponibilidade de grandes quantidades de dados e a necessidade iminente de transformar esses dados em informações e conhecimentos úteis. As informações e o conhecimento adquiridos podem ser usados para aplicações que vão desde análise de mercado, detecção de fraudes e retenção de clientes, até controle de produção e exploração científica (HAN e KAMBER, 2006, p.3).

Segundo Baker, Isotani e Carvalho (2011), para o contexto educacional, tem-se a Mineração de Dados Educacionais (do inglês, "*Educational Data Mining*", ou EDM).

> A EDM é definida como a área de pesquisa que tem como principal foco o desenvolvimento de métodos para explorar conjuntos de dados coletados em ambientes educacionais. Assim, é possível compreender

> de forma mais eficaz e adequada os alunos, como eles aprendem, o papel do contexto na qual a aprendizagem ocorre, além de outros fatores que influenciam a aprendizagem (BAKER; ISOTANI; CARVALHO, 2011).

As bases de dados são mantenedoras de um conjunto de dados relacionados e grandes quantidades de informações fundamentais para o negócio. A fim de obter informações mais específicas relacionadas ao desempenho dos alunos com base nos indicadores educacionais, notou-se a viabilidade e a necessidade de realizar análises utilizando técnicas de mineração de dados nas bases de dados dos cursos superiores do Instituto Federal de Educação, Ciência e Tecnologia de Minas Gerais *campus* São João Evangelista (IFMG-SJE) e identificar possíveis padrões.

Algumas perguntas foram elaboradas e serão respondidas com base em todas as análises realizadas no decorrer deste trabalho, dessa forma pode-se entender algumas causas que impactam ou não para a desistência, permanência e conclusão de um aluno.

a) A quantidade de reprovação nas disciplinas impacta os indicadores de fluxo?
b) A quantidade de faltas nas disciplinas impacta os indicadores de fluxo?
c) Qual gênero é mais propenso a desistir do curso?
d) A pandemia de alguma forma afetou os indicadores de fluxo?
e) Os dados socioeconômicos influenciam nos indicadores?

1.1 Objetivo Geral

O objetivo geral deste trabalho é identificar padrões relacionados aos indicadores de fluxo dos alunos a partir da análise das bases de dados do IFMG-SJE.

1.2 Objetivos específicos

Os objetivos específicos são:

- Aplicar técnicas de mineração de dados nas bases do IFMG-SJE;
- Identificar padrões a partir da geração de regras de associação ou agrupamento;

- Utilizar os indicadores de fluxo para nortear todas as análises a serem realizadas.

1.3 Justificativa

Apesar das aplicações de Mineração de Dados serem implementadas nos mais diversos setores, como saúde, varejo, indústria, serviços financeiros, entre outros, o foco deste trabalho é o setor de educação. O conhecimento contido nos grandes conjuntos de dados possui informações relevantes para treinamento de modelos computacionais. “Esse conhecimento poderá ser usado para a tomada de decisão estratégica, como controle de processos, gestão da informação e conhecimento, processamento de consultas e muitas outras aplicações.” (CASTRO; FERRARI, 2016).

O foco principal deste trabalho é identificar a *persona*, ou seja, as características dos alunos de acordo com os indicadores de fluxo (desistência, permanência e conclusão). Essas análises servirão como base para que o IFMG-SJE tome ações para minimizar os impactos da evasão. O trabalho tem ênfase na análise de dados das bases internas do IFMG-SJE, utilizando a linguagem Python.

1.4 Contribuições

O trabalho tem como principal contribuição identificar padrões relacionados ao desempenho dos alunos que fazem ou fizeram parte da Instituição. Com um incentivo a mais que ratifica a importância deste trabalho, observou-se que este é pioneiro a abordar a análise dos dados tendo como premissa o desempenho educacional na instituição em questão, o que pode incentivar outros alunos a explorarem mais o tema, acerca dos dados e das análises que serão geradas.

1.5 Estrutura do trabalho

O trabalho está organizado da seguinte forma: o capítulo dois aborda o referencial teórico, tendo como objetivo verificar o estado do problema a ser pesquisado, sob o aspecto teórico e de outros estudos e pesquisas já realizadas. No capítulo três, a metodologia utilizada para executar o processo de pesquisa do trabalho. No capítulo quatro, mostram-se os

resultados e discussões acerca das análises realizadas, por fim o capítulo cinco apresenta as considerações finais do trabalho.

2 REFERENCIAL TEÓRICO

Este estudo procurou entender através da literatura disponível os indicadores de desempenho educacional no âmbito superior, o processo de mineração de dados e como o processo de descoberta do conhecimento, conhecido como KDD - *Knowledge Discovery In Databases*, se aplica e molda este trabalho científico.

2.1 Indicadores de desempenho educacional

De acordo com a Associação Brasileira de Mantenedoras de Ensino Superior (ABMES), em 20 de dezembro de 1996, foi promulgada a Lei nº 9.394, de Diretrizes e Bases da Educação Nacional (LDB), com base nessa lei a educação superior ingressou numa fase que passou a exigir o avanço da profissionalização das ações acadêmicas e das ações de gestão desenvolvidas pelos profissionais das instituições de educação superior públicas e privadas, visando consolidar no setor educacional o debate da qualidade.

De acordo com a Lei nº 9.394, de 20 de dezembro de 1996:

> Art. 1°. A educação abrange os processos formativos que se desenvolvem na vida familiar, na convivência humana, no trabalho, nas instituições de ensino e pesquisa, nos movimentos sociais e organizações da sociedade civil e nas manifestações culturais.[...]
> Art. 9°. A União incumbir-se-á de:
> I – elaborar o Plano Nacional de Educação, em colaboração com os Estados, o Distrito Federal e os Municípios;[...] (BRASIL, 1996).

Para Cardoso (2004), ratificado por Cunha e Carrilho (2005), o ensino superior ao longo das últimas décadas vem sofrendo com as acentuadas mudanças da sociedade. Neste sentido a universidade necessita de uma nova organização, englobando e ressignificando a maneira da sociedade produzir, criando e difundindo seus valores de forma a promover a melhoria da condição humana em suas múltiplas dimensões. Para tanto é necessário que a universidade reveja seus métodos, suas práticas, objetivos, currículo e até metodologias de aprendizagem.

De acordo com Tam (2001), citado por ABMES (2019), todas as IES possuem a responsabilidade institucional de melhoria contínua na qualidade dos serviços prestados, com o objetivo de alcançar metas de aperfeiçoamento que podem ser expressas por meio de um sistema de avaliação com base em valores quantitativos, denominados *Performance Indicators* (PIs), os quais serão abordados e explicados posteriormente. Independentemente de qual seja a organização, os PIs exercem papel importante na avaliação externa do seu funcionamento, sendo de grande importância para o controle da qualidade

Marchelli (2007), afirma que ao analisar a complexidade do processo de ensino e aprendizagem, deve-se levar em conta que os PIs não constituem os únicos parâmetros a serem considerados, por não perceberem variações subjetivas e de caráter individual e único de cada IES. Logo, a utilidade dos indicadores de desempenho como instrumentos centrais para o controle da qualidade é bastante questionável, mas, por sua objetividade e precisão, eles têm um papel de destaque no conjunto dos procedimentos adotados.

Segundo o Ministério da Educação, o Censo da Educação Superior é o principal instrumento de pesquisa de avaliação das IES visando a qualidade do ensino no Brasil, sendo realizado anualmente pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). O Censo utiliza as informações do cadastro do Sistema e-MEC, em que são mantidos os registros de todas as IES, seus cursos e locais de oferta. A partir desses registros, são coletadas informações sobre a infraestrutura das IES, vagas oferecidas, candidatos, matrículas, ingressantes, concluintes e docentes, nas diferentes formas de organização acadêmica e categoria administrativa.

O objetivo da coleta, de acordo com o Ministério da Educação (2020a), é oferecer informações estatísticas confiáveis, que permitam conhecer e acompanhar o sistema brasileiro de educação superior. Além disso, auxilia o Ministério da Educação por meio de informações estatísticas, possibilitando um êxito maior nas atividades de acompanhamento e avaliação, programas de expansão e de melhoria da qualidade deste nível de ensino, entre outros.

> Os dados serão utilizados para calcular indicadores que fundamentam a formulação e a implementação de políticas públicas e contribuir com o trabalho dos gestores das IES e demais gestores de governo, de instituições de âmbito público ou privado, pesquisadores, especialistas e estudantes do Brasil e de outros países, bem como de organismos internacionais (MINISTÉRIO DA EDUCAÇÃO, 2020a).

Segundo o Ministério da Educação (2020a), o INEP possui 3 grandes indicadores de avaliação para o ensino superior, para avaliar desde o ingresso do discente no curso até a conclusão ou desistência. Esses indicadores estão divididos em:

I. Permanência no curso de ingresso, onde é feito uma análise com base nos discentes que possuem vínculo ativo com seu curso de ingresso em um determinado ano de referência.
II. Desistência do curso de ingresso, indicador que traz as desistências (desvinculações) ou transferências do curso de ingresso em um determinado ano de referência.
III. Conclusão no curso de ingresso, referente aos discentes que se formam em um determinado ano de referência.

Para ABMES (2019), a qualidade na educação deve ser um norte para o desenvolvimento da sociedade e dos seus indivíduos. Só será possível se seguirmos esse princípio norteador para criarmos formas de avaliação que permitam acessos às instituições de ensino superior, que possibilite a permanência do estudante na instituição.

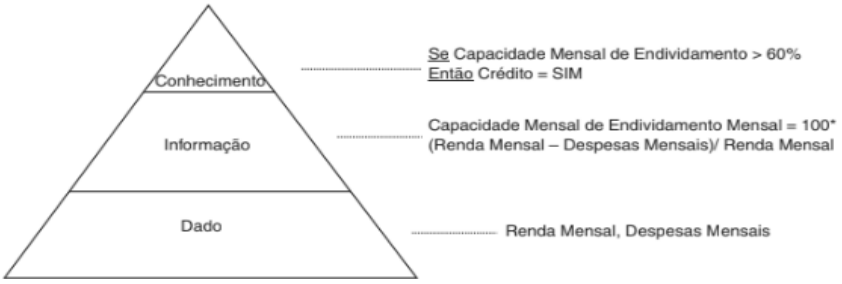
Ribeiro (2015) citado por ABMES (2019) apresenta uma possibilidade de classificação dos tipos de avaliação, conforme a finalidade, dividindo-a em dois grupos:

I. Modelos educativos, ou formativos, cuja principal finalidade é desenvolver e aprimorar a qualidade do trabalho produzido pela instituição avaliada. Esse modelo é caracterizado pela ênfase na análise qualitativa e incentiva o envolvimento de todos os segmentos da instituição na construção e execução do processo;
II. Os modelos regulatórios, cuja principal finalidade é garantir o cumprimento das regras de funcionamento preestabelecidas para o sistema, garantindo, o nível de qualidade do trabalho das instituições avaliadas. O modelo tem como principal característica a ênfase na análise quantitativa.

2.2 *KDD - Knowledge Discovery In Databases*

Antes de introduzir os conceitos referentes ao KDD e a Mineração de Dados respectivamente, a fim de proporcionar um melhor entendimento é importante destacar as diferenças e a hierarquia entre dado, informação e conhecimento, conforme ilustra a Figura 1.

Figura 1. Hierarquia entre dados, informação e conhecimento



Fonte: GOLDSCHMIDT; PASSOS, 2005, p. 2

Na base da pirâmide encontram-se os dados brutos/primários, adquiridos e armazenados por recursos da Tecnologia da Informação e Comunicação (TICs). Na camada de Informação, encontram-se os dados processados, essa camada possui definições e contextos claros, são utilizados muitos recursos da Tecnologia da Informação para processar os dados e obter as informações necessárias.

O conhecimento está no topo da pirâmide, podendo ser também um padrão, ou um conjunto de padrões, onde os dados e informações podem ser relacionados e inclusos. Geralmente, o conhecimento não pode ser obtido nas bases de dados utilizando recursos tradicionais.

De acordo com Goldschmidt e Passos (2005), em 1989 foi formalizado o termo KDD, em alusão à ideia de procurar conhecimento através das bases de dados. Em 1996 um grupo de pesquisadores sugeriu o que é uma das definições mais utilizadas:

> “KDD é um processo, de várias etapas, não trivial, interativo e iterativo, para identificação de padrões compreensíveis, válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados” (FAYYAD et al., 1996a apud GOLDSCHMIDT; PASSOS, 2005).

Fayyad, Piatetsky-Shapiro e Smyth (1996) apontam que o processo KDD abrange qualquer seleção necessária utilizando o banco de dados, o pré-processamento, a subamostragem e transformações; aplicar métodos de mineração de dados (algoritmos) para listar os padrões e; avaliar o que foi obtido através de mineração de dados, para então abstrair o conhecimento.

Goldschmidt (2003), tendo como propósito melhor situar a área de KDD, apresenta uma taxonomia vista na Figura 2, das atividades na área da Descoberta de

Conhecimento em Bases de Dados. Essa taxonomia mostra a diversidade de atividades relacionadas ao contexto de KDD.

Figura 2. Taxonomia de Atividades na Área de KDD.

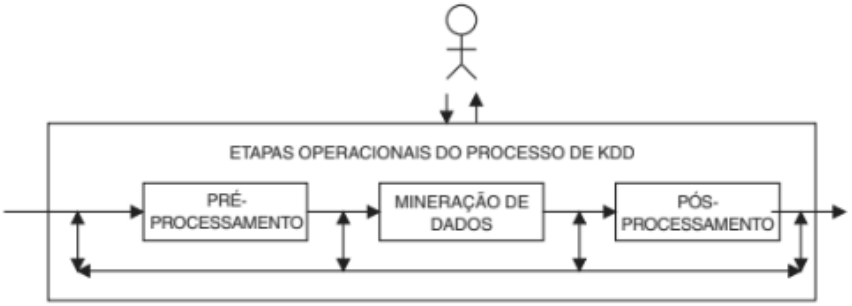
Fonte: GOLDSCHMIDT; PASSOS, 2005, p. 6

Segundo Goldschmidt e Passos (2005), as atividades na área de KDD podem ser organizadas em três grandes grupos:

I. Desenvolvimento Tecnológico: engloba todas as iniciativas que podem ser utilizadas na busca por novos conhecimentos nas grandes bases de dados.
II. Execução de KDD: abrange toda a parte de busca real por conhecimentos nas bases de dados.
III. Aplicação de Resultados: a partir dos modelos úteis de conhecimento adquiridos através das grandes bases de dados, as atividades retornam para a parte de aplicar os resultados no contexto que foi realizado o processo de KDD. O desenvolvimento de sistemas que utilizem conhecimentos extraídos de bases de dados tem propiciado valiosas ferramentas de apoio à decisão.

Goldschmidt e Passos (2005), afirmam que são necessárias várias etapas operacionais para obter a Descoberta de Conhecimento em Bases de Dados, a Figura 3 ilustra essas etapas. Resumidamente, na etapa de pré-processamento existem funções voltadas à captação, à organização e ao tratamento dos dados.

Figura 3. Etapas Operacionais do Processo de KDD



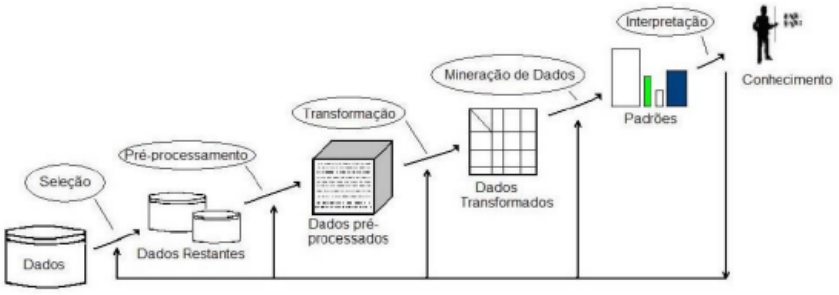
Fonte: GOLDSCHMIDT; PASSOS, 2005, p. 3

Ainda de acordo com Goldschmidt e Passos (2005), na etapa de pré-processamento, o objetivo é a preparação dos dados para a etapa seguinte, a etapa de Mineração de Dados. Durante a Mineração de Dados, são feitas buscas por conhecimentos úteis no ambiente da aplicação de KDD.

Para Fayyad et al. (1996a), na etapa de pós-processamento, ocorre o tratamento do conhecimento adquirido com a mineração de dados. Nem sempre é necessário realizar essa etapa, mas, seu objetivo é viabilizar a avaliação da utilidade do conhecimento descoberto.

Fayyad et al. (1996) abordam o processo KDD sendo constituído de várias etapas, executadas de forma interativa e iterativa. São interativas porque envolvem a cooperação da pessoa responsável pela análise de dados, cujo conhecimento sobre o domínio orientará a execução do processo. Já a iteração deve-se ao fato de que esse processo não é executado de forma sequencial, mas envolve repetidas seleções de parâmetros e conjunto de dados, bem como a aplicação das técnicas de *Data Mining* e a análise dos resultados obtidos, a fim de refinar os conhecimentos extraídos. O processo de KDD é composto por cinco fases, que podem ser vistas na Figura 4 e são nomeadas como: seleção de dados, pré-processamento, transformação, mineração e interpretação/avaliação.

Figura 4. Fases da descoberta de conhecimento em bases de dados



Fonte: CASTANHEIRA[1], 2008 apud CORNELIUS JUNIOR, 2015, p. 19

2.3 Pré-processamento

A etapa de Pré-processamento, de acordo com Goldschmidt e Passos (2005), concebe todas as partes relacionadas à captação, organização e tratamento dos dados. É uma parte muito importante para o processo de descoberta do conhecimento. Nessa etapa, o objetivo é a preparação dos dados para os algoritmos de Mineração de Dados. Esse ciclo abrange desde a correção de dados errados até o ajuste da formatação dos dados para os algoritmos de Mineração de Dados.

2.3.1 Seleção de dados

A seleção dos dados segundo Castro e Ferrari (2016), compreende, em essência, a identificação de quais informações, dentre as bases de dados existentes, devem ser efetivamente consideradas durante o processo de KDD.

> [...] Selecionar um conjunto de dados ou focar em um subconjunto de variáveis ou amostras de dados, no qual a descoberta é a ser executada (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996, p.42, tradução nossa).[2]

[1] CASTANHEIRA, Luciana Gomes. Aplicação de Técnicas de Mineração de Dados em Problemas de Classificação de Padrões. Belo Horizonte: UFMG, 2008.
[2] No original: [...] Selecting a data set, or focusing on a subset of variables or data samples, on which discovery is to be performed.

De acordo com Witten e Frank (2005), a redução de dados seja de forma manual ou automatizada, melhora o desempenho dos algoritmos. Nesse estágio, o foco é selecionar os atributos mais relevantes e excluir os inadequados para que, com essa diminuição de dimensionalidade, se consiga uma representação compacta e de fácil interpretação dos dados pertinentes.

2.3.2 Pré-processamento e limpeza de dados

Conforme Goldschmidt e Passos (2005), essa etapa compreende as funções relacionadas à captação, à organização, ao tratamento e à preparação dos dados para a etapa da Mineração de Dados. Essa etapa possui fundamental relevância no processo de descoberta de conhecimento, pois busca compreender desde a correção de dados errados até o ajuste da formatação dos dados para os algoritmos de Mineração de Dados a serem utilizados.

De acordo com Fayyad, Piatetsky-Shapiro e Smyth (1996), ratificado por Witten e Frank (2005), os dados devem ser pré-processados, para que haja uma seleção de um subconjunto que será utilizado no modelo de aprendizado. O desempenho dos modelos podem ser melhorados com algumas operações básicas que devem ser executadas com os dados selecionados.

> A fase de limpeza dos dados envolve uma verificação na consistência das informações, a correção de possíveis erros e o preenchimento ou a eliminação de valores desconhecidos e redundantes, além da eliminação de valores não pertencentes ao domínio. A execução dessa fase tem como objetivo, portanto, corrigir a base de dados, eliminando consultas desnecessárias que poderiam ser executadas futuramente pelos algoritmos de Mineração de Dados, afetando o desempenho destes algoritmos (GOLDSCHMIDT; PASSOS, 2005, p.37).

2.3.3 Transformação dos dados

A etapa de transformação ou formatação dos dados analisa os dados obtidos na etapa de pré-processamento e os reorganiza de uma forma específica para que possam ser interpretados na etapa seguinte. Han et al (2012) afirma que nessa etapa os dados são

transformados ou consolidados para que o processo de mineração resultante possa ser mais eficiente, e os padrões encontrados possam ser mais fáceis de entender.

Os métodos de transformação de dados visam modificar ou consolidar os dados em formas apropriadas aos processos de mineração. Segundo Goldschmidt e Passos (2005), esses métodos podem ser classificados em: padronização e normalização. A padronização tem como objetivo principal resolver as diferenças de unidades e escalas dos dados, já a normalização é um processo de transformação dos dados que objetiva torná-los mais apropriados à aplicação de algum algoritmo de mineração.

2.4 Mineração de dados

De acordo com Fayyad, Piatetsky-Shapiro e Smyth (1996), a Mineração de Dados é uma das etapas do KDD, nessa etapa os métodos são realizados através de repetição e interação para a obter padrões. Os objetivos das tarefas dirão quais os tipos de algoritmos serão utilizados. O processo de Mineração de Dados pode ser feito de forma direta ou indireta:

- Mineração de Dados Direta: um modelo é gerado a partir de uma base de dados, e é utilizado no processo de mineração de uma determinada entrada de dados (registros a serem classificados). Ao final deste processo, são geradas informações úteis para o usuário e para as tarefas (classe de cada registro).
- Mineração de Dados Indireta: minera-se a informação a partir da base de dados, dando origem a padrões, que são interpretados a fim de gerar informações úteis, isto é, informações que levem benefício ao usuário ou às tarefas.

2.4.1 Tarefas e técnicas de mineração de dados

Silva e Ferrari (2016, p. 50) apontam que as tarefas de mineração de dados são métodos para extrair conhecimento em bases de dados. Essas tarefas podem ser classificadas em duas categorias, as tarefas descritivas que caracterizam as propriedades gerais dos dados; e tarefas preditivas que fazem inferência a partir dos dados objetivando predições. A aplicação

de técnicas de mineração de dados possibilita extrair informações escondidas nos dados e informações úteis e indispensáveis para a tomada de decisão estratégica.

Patrício Júnior (2012) divide as tarefas preditivas em classificação e regressão e as tarefas descritivas mais conhecidas são as regras de associação, clusterização e sumarização.

> Os algoritmos de classificação possuem o objetivo de "aprender"a classificar registros em classes pré definidas. As tarefas de regressão são utilizadas quando o objetivo é prever resultados futuros a partir de registros antigos. Os algoritmos de agrupamento procuram identificar grupos ou clusters em um conjunto de registros a partir de seus atributos. Já quando se deseja procurar relacionamentos entre os registros de um BD, utiliza-se os algoritmos de Associação (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

2.4.2 Análise preditiva dos dados

Witten, Frank e Hall (2011), abordam a predição como um dos objetivos fundamentais da mineração de dados, utilizando algumas variáveis presentes no banco de dados, com a finalidade de prever valores desconhecidos ou futuros de outras variáveis que sejam de interesse.

De acordo com Silva e Ferrari (2016, p. 303), busca-se um modelo, que a partir de conhecimentos de casos anteriores, possa permitir fazer uma previsão de valores de determinados atributos em novas circunstâncias. Nas tarefas preditivas o desempenho de um algoritmo é avaliado usando taxas de erro, como o erro de classificação ou de estimação. Como a base de dados de treinamento é uma amostra de uma base de dados, e ela está sujeita a ruídos e impurezas, o erro quando encontrado durante o treinamento é uma estimativa do erro real.

2.4.2.1 Classificação

A classificação é o processo de encontrar um modelo (ou função) que descreve e distingue orienta classes de dados ou conceitos, com a finalidade de poder usar o modelo para

prever a classe de objetos cujo rótulo de classe é desconhecido. (Han; Kamber, 2006, p.53, tradução livre).

Para Goldschmidt e Passos (2005), a tarefa de classificação é considerada uma das tarefas de KDD mais importantes e mais populares. Pode ser compreendida como a busca por uma função que permita associar corretamente cada registro Xi de um banco de dados a um único rótulo categórico, Yj, denominado classe. Uma vez identificada, essa função pode ser aplicada a novos registros de forma a prever a classe em que tais registros se enquadram.

2.4.3 Análise descritiva dos dados

A descrição também é um dos objetivos fundamentais da mineração de dados, busca por padrões que descrevem os dados, de forma que possam ser interpretáveis pelos usuários, a fim de encontrar respostas que confirmem ou neguem as hipóteses (WITTEN; FRANK; HALL, 2011).

> A análise descritiva dos dados é uma etapa inicial do processo de mineração que não requer elevado nível de sofisticação, utiliza-se ferramentas capazes de medir, explorar e descrever características intrínsecas aos dados.Nesse caso busca-se por um modelo que descreva, de forma compreensível pelo homem, o conhecimento existente em um conjunto de dados. (SILVA; FERRARI, 2016).

2.4.3.1 Regras de associação

Para Witten e Frank (2005, tradução livre)[3], as regras de associação diferenciam-se das regras de classificação, elas podem prever qualquer atributo e suas combinações, não apenas a classe. Além disso, as regras de associação não se destinam para serem usadas como um conjunto, como são as regras de classificação.

A tarefa de regras de associação consiste em encontrar conjuntos de itens que ocorram simultaneamente e de forma frequente em um banco de dados:

[3] No original: Association rules are really no different from classification rules except that they can predict any attribute, not just the class, and this gives them the freedom to predict combinations of attributes too. Also, association rules are not intended to be used together as a set, as classification rules are.

> As regras de associação se diferenciam marcantemente dos outros algoritmos, pois enfatizam a análise das relações entre os atributos, e não entre os objetos da base. Essa regra busca encontrar padrões do tipo X→Y, o quanto X implica em Y onde X e Y são conjuntos distintos, ou seja, registros da base de dados que satisfazem à condição em X também satisfazem à condição em Y (SILVA; FERRARI, 2016).

Santos (2017) evidencia que as regras de associação encontram-se entre um dos mais importantes tipos de conhecimento que podem ser minerados em bases de dados. Estas regras representam padrões de relacionamento entre itens de uma base de dados.

2.4.3.2 Clusterização

Para Silva e Ferrari (2016), o agrupamento (*clustering*) é o processo de separar um conjunto de objetos em grupos (do inglês *clusters*) de itens semelhantes. Normalmente utilizado para identificar os grupos, onde cada grupo formado pode ser visto como uma classe de objetos. Como não há rotulação nas classes de treinamento, logo não são *apriori*, o processo então é nomeado como aprendizagem não supervisionada.

> A tarefa de Agrupamento, também denominada de clusterização ou segmentação, é utilizada para dividir os dados em grupos (*clusters*). O objetivo é que os objetos dentro de um grupo sejam semelhantes e diferentes de outros objetos de outros grupos. Quanto maior a semelhança dentro de um grupo e maior a diferença entre grupos, melhor ou mais distinto será o agrupamento (TAN, STEINBACH, KUMAR, 2006).

Goldschmidt e Passos (2005) apontam que o processo de clusterização requer ao usuário determinar qual o número de grupos a serem considerados. Com base neste número, os registros de dados são então separados nos grupos de forma que registros similares fiquem nos mesmos grupos e registros diferentes em grupos distintos. Uma vez tendo esses grupos, é

possível fazer uma análise dos elementos que compõem cada um deles, identificando as características comuns aos seus elementos e, desta forma, podendo criar um rótulo que representa cada grupo.

2.4.3.3 Sumarização

Goldschmidt; Passos, 2005, afirmam que a sumarização é uma tarefa muito comum em KDD, e tem como objetivo procurar identificar e indicar características comuns entre conjuntos de dados.

Técnicas descritivas de sumarização de dados podem ser usadas para identificar as propriedades típicas de seus dados e destacar quais valores de dados devem ser tratados como ruído ou *outlier* (HAN; KAMBER, 2006, tradução livre)[4].

Para Goldschmidt e Passos (2005), a sumarização consiste em identificar e apresentar agrupamentos de objetos, de forma concisa e compreensível, e as principais características de conjunto de dados. Além disso, busca gerar descrições para caracterização resumida dos dados, também chamadas de descrições de classes, e a discriminação entre eles. Através da descrição de classes tem-se a descrição de conceitos, que pode ser interpretada como uma generalização dos dados, das características mais relevantes dentre os registros analisados.

2.4.3.4 Detecção de desvios

De acordo com Han e Kamber (2001) ratificado por Weis e Indurkhya (1999), citado por da Costa Côrtes, Porcaro e Lifshitz (2002), esta funcionalidade objetiva encontrar conjuntos de dados que não obedecem ao comportamento ou modelo de dados. Uma vez encontrados podem ser tratados ou descartados para utilização no processo de mineração de dados. Trata-se de uma importante avaliação nos dados no sentido de descobrir probabilidades crescentes de desvio ou riscos associados aos vários objetivos traçados inicialmente na mineração dos dados. Detectar esses desvios é muito análogo às técnicas utilizadas nas análises estatísticas, onde são aplicados testes de significância que assumem uma distribuição, utilizando medidas estatísticas do tipo média aritmética e desvio padrão para aferir essas diferenças.

---

[4] No original: Descriptive data summarization techniques can be used to identify the typical properties of your data and highlight which data values should be treated as noise or outlier.

Goldschmidt e Passos (2005) afirmam que, enquanto a repetição de padrões em outras tarefas do KDD é a principal característica para a busca do conhecimento, na detecção de desvios o objetivo é identificar padrões que aparecem com pouca frequência e que são diferentes dos valores comumente registrados.

2.4.3.5 Clusterização com Classificação

Para Goldschmidt e Passos (2005), quando há a junção de clusterização e classificação elas se tornam uma tarefa composta, que é muito comum em aplicações de KDD. É aplicado quando os dados não estão enquadrados em classes predefinidas. Essa tarefa é composta por dois tópicos basicamente:

a) Reunir os dados que são similares. Quando é feita a clusterização, geralmente os métodos incluem mais um atributo no conjunto de dados original para referenciar à qual cluster cada registro pertence.

b) Os rótulos de *cluster* agora são tratados como classes, e cada registro do novo conjunto de dados está sendo delimitado em alguma classe. Quando ocorre esse cenário, os algoritmos de classificação podem ser aplicados. A partir de então, algoritmos de classificação podem ser aplicados de forma a gerar elos de conhecimento que possam prever a classificação de novos registros a partir das características dos dados (GOLDSCHMIDT;PASSOS, 2005).

2.4.3.6 Clusterização com Sumarização

Goldschmidt e Passos (2005) falam sobre a tarefa de "Clusterização Sumarização", é um tipo de tarefa composta e muito comum em aplicações de KDD, ela encadeia as tarefas primárias de cada um das técnicas. É aplicável quando o conjunto de dados tem pouca semelhança entre seus dados, esta tarefa consiste em:

a) Reunir os dados que são similares utilizando algum método de clusterização dos dados. Quando o método é aplicado o conjunto de dados original é segmentado para a mesma quantidade de *clusters* gerados.

b) O algoritmo de sumarização vai descrever os novos conjuntos de dados, dizendo suas principais características dos registros de cada conjunto.

2.5 Pós processamento

De acordo com Lima (2019) as técnicas utilizadas para representar os dados processados, de forma visual, facilitam a compreensão humana, pois desse modo se associa melhor as informações adquiridas. Entretanto, chegar nessa etapa final não acaba com a tomada de conhecimento, já que deve ser tudo analisado por especialistas da área em relação à validade dos resultados adquiridos e na qualidade dos dados gerados.

> Essa fase envolve a visualização, a análise e a interpretação do modelo de conhecimento gerado pela etapa de Mineração de Dados. Em geral, é nesta etapa que ocorre a avaliação dos resultados obtidos e definem-se novas alternativas de investigação dos dados (GOLDSCHMIDT; PASSOS, 2005).

2.5.1 Interpretação

A etapa de interpretação segundo Bernardini (2017), é onde ocorre a interpretação dos padrões minerados com o possível regresso a uma das fases anteriores para maior interação ou documentação. O conhecimento descoberto será consolidado e incorporado ao sistema ou elaboração de relatórios para as partes interessadas, também é possível verificar e resolver potenciais conflitos com conhecimento tido como verdadeiro ou previamente extraído.

2.6 Métodos e procedimentos

2.6.1 Levantamento inicial

O primeiro momento envolve a definição sobre “o que fazer” diante da base de dados a ser analisada. Neste momento, devem ser executadas as etapas de “Levantamento Inicial” e de “Definição de Objetivos”. O Levantamento Inicial compreende um exame preliminar da base de dados, procurando obter informações sobre a natureza e o propósito dos dados a serem analisados.

Existem algumas considerações técnicas que devem ser verificadas no início de um processo de KDD:

I. Identificar as pessoas e áreas envolvidas no processo de KDD.
II. Realizar um levantamento do hardware e software existentes.
III. Fazer um inventário das bases de dados disponíveis, tanto bases de dados internas quanto externas.
IV. Verificar a existência de *Data Warehouses*.
V. Compreender o significado e perceber a relevância dos atributos disponíveis. Metadados acerca das bases de dados e seus atributos devem ser documentados.
VI. Procurar identificar que critérios podem ser adotados para mensurar o sucesso do processo.
VII. Avaliar a qualidade dos dados disponíveis. Deve-se procurar identificar o propósito para o qual os dados foram coletados, assim como a caracterização do nível de ruído envolvido. Bases de dados poluídas requerem o uso de ferramentas adequadas ao processo de limpeza.
VIII. Verificar se os dados estão disponíveis em quantidade suficiente para o processo de KDD. Bases de dados pequenas ou que contenham dados pouco representativos das condições normais do domínio da aplicação podem inviabilizar o processo de KDD.

2.6.2 Definição de objetivos

Na definição de objetivos, devem ser identificadas quais tarefas de Mineração de Dados são viáveis. Neste momento, devem ser formulados alguns requisitos quanto ao modelo de conhecimento a ser produzido. Nessa etapa mais de um objetivo pode ser constituído.

A definição dos objetivos em qualquer aplicação de KDD requer, primeiramente, um entendimento claro da situação vigente no ambiente onde será realizado o processo. Esse entendimento começa a ser formado desde a etapa de levantamento inicial de informações. Uma análise da natureza dos dados pode fornecer alguns indicadores de possíveis tarefas de mineração de Dados, que devem ser consideradas na formulação dos objetivos do processo de KDD.

2.6.3 Planejamento de atividades

A partir da escolha de um objetivo, a abordagem proposta é direcionada para a definição sobre “como fazer”, correspondendo à etapa de “Planejamento de Atividades”. Nesta etapa devem ser definidas as alternativas associadas ao objetivo escolhido, um plano de ação é uma sequência válida de métodos de KDD. Essa metodologia sugere que os planos de ação sejam constituídos a partir de cada método de mineração de dados aplicável à tarefa de KDD associada ao objetivo selecionado.

Neste momento, deve-se planejar quais métodos de pré-processamento devem ser utilizados, incluindo a ordem de aplicação. Por exemplo, suponha que a base de dados possua dados não normalizados e que o método candidato tenha como precondição que todos os dados estejam normalizados.

2.6.4 Execução dos planos de ação

Para cada plano de ação selecionado, iniciam-se os trabalhos de execução ordenada das ações previstas no plano. A etapa de “Execução dos Planos de Ação” corresponde à aplicação propriamente dita dos métodos de KDD. Neste momento podem e devem ser experimentados de forma coerente diversos valores nos parâmetros dos algoritmos envolvidos.

A execução de um plano de ação compreende a execução ordenada dos métodos que compõem o plano, sendo essa execução feita em ciclos. A cada ciclo, deve-se executar total ou parcialmente o plano de ação, procurando obter melhores resultados. Em geral, esta etapa ocorre de forma integrada à etapa de “Avaliação de Resultados”.

2.6.5 Avaliação de resultados

Finalmente, a abordagem proposta é concluída pela etapa de “Avaliação de Resultados", que corresponde à "análise do que foi feito”. Em geral, a avaliação de resultados pode ser realizada de forma mais efetiva após a execução dos métodos de mineração de dados. Neste momento, as características do modelo de conhecimento gerado devem ser

confrontadas com as expectativas quanto ao modelo formuladas na etapa de “Definição de Objetivos”.

2.7 Python

A linguagem Python foi desenvolvida por Guido van Rossum, no final dos anos 80, na Holanda, a partir de uma linguagem existente nessa época, chamada ABC. Essa linguagem tinha como principal foco os físicos e engenheiros.

> O Python possui uma sintaxe clara e concisa, que favorece a legibilidade do código fonte, tornando a linguagem mais produtiva. A linguagem inclui diversas estruturas de alto nível (listas, dicionários, data / hora, complexos e outras) e uma vasta coleção de módulos prontos para uso, além de frameworks de terceiros que podem ser adicionados. Também possui recursos encontrados em outras linguagens modernas, tais como: geradores, introspecção, persistência, metaclasses e unidades de teste (BORGES, 2010).

O Python possui várias características interessantes para a comunidade científica, ela é multiplataforma; os aplicativos que são desenvolvidos em Python, em uma plataforma podem ser compartilhados com outras plataformas; além de ser um software livre e orientado a objetos.

2.8 Trabalhos correlatos

A mineração de dados tornou-se campo de pesquisa e objeto de estudo pela necessidade de extrair conhecimento de conjuntos de dados através de um processo. Na literatura foram encontrados diversos trabalhos relacionados ao tema proposto, realçando a importância da mineração de dados na detecção de irregularidades e nas possibilidades de demonstrar os *gaps* existentes no desempenho dos alunos em nível superior. A seguir são discutidos trabalhos baseados em mineração de dados no contexto educacional:

I. OLIVEIRA (2013) descreve um trabalho cujo objetivo é buscar informações implícitas na base de dados dos candidatos inscritos nos processos seletivos do

Instituto Federal de Minas Gerais – Campus São João Evangelista, através da aplicação do processo de Descoberta de Conhecimento em Base de Dados (DCBD), aplicando-se técnicas de Regra de Associação, Árvore de Decisão e Agrupamento. Além de ser um estudo realizado na mesma instituição ao qual esse trabalho se refere, pode-se perceber semelhanças principalmente no uso da mineração de dados e algumas técnicas. Os trabalhos se diferenciam na utilização do DCBD e algoritmos de mineração de dados.

II. CORNELIUS JUNIOR (2015) descreve um trabalho cujo objetivo é utilizar a mineração de dados para identificar alunos com perfil de evasão do ensino superior, utilizando dados cedidos pela Universidade de Santa Cruz do Sul – UNISC. A similaridade do trabalho está primeiramente no uso da mineração de dados no contexto educacional, além disso o autor utilizou técnicas de mineração conhecidas como classificação e associação, para a realização de experimentos e por fim chegou-se a conclusão dos principais fatores que contribuem para a evasão de alunos. O trabalho de Cornélius difere-se pois possui foco principal no indicador de evasão, já neste trabalho foca nos três indicadores de fluxo.

III. SANTOS (2017) descreve um trabalho cujo objetivo é realizar uma análise de dados para uma avaliação de desempenho dos alunos do instituto de computação da UFF. O trabalho se assemelha a este em questão, pois tem como objetivo empregar técnicas de análise de dados, em especial, técnicas de mineração, junto a todo o processo de KDD, utilizando dados históricos dos alunos dos cursos de graduação, ligados ao instituto, sendo possível realizar as análises e obter informações. A principal diferença está no uso de algoritmos de mineração, por exemplo a árvore de decisão, não utilizados no trabalho desenvolvido.

IV. FONSACA e DOROCINSKI (2018) descreve um trabalho com o uso da mineração de dados para o auxílio à tomada de decisão da UFPR, este projeto objetiva propor uma solução em formato de sistema para auxílio ao combate dos fenômenos da evasão e retenção escolar através da prevenção, a qual se dá na identificação antecipada dos possíveis casos, o sistema apresentado também possui funcionalidades que ajudam professores e equipes que trabalham na universidade a gerenciar casos de jubilamento e atendimento à alunos com problemas de desempenho. A principal diferença está na proposta de desenvolver um sistema, o que não foi proposto no trabalho desenvolvido.

3 METODOLOGIA

Entende-se Metodologia como o estudo do método para se buscar determinado conhecimento. Trata-se das formas de se fazer ciência e trata-se dos procedimentos, ferramentas e caminhos. Bruyne (1991) ratificado por Marconi e Lakatos (2017) afirma que:

> A metodologia deve ajudar a explicar não apenas os produtos da investigação científica, mas principalmente seu próprio processo, pois suas exigências não são de submissão estrita a procedimentos rígidos, mas antes da fecundidade na produção dos resultados.(BRUYNE, 1991 p. 29).

3.1 Método científico

O método de pesquisa é um conjunto de procedimentos e técnicas utilizados para coletar e analisar os dados, fornecendo os meios para se alcançar o objetivo proposto, Marconi e Lakatos (2017) evidenciam o método como:

> Método é o conjunto das atividades sistemáticas e racionais que, com maior segurança e economia, permite alcançar o objetivo de produzir conhecimentos válidos e verdadeiros, traçando o caminho a ser seguido, detectando erros e auxiliando as decisões do cientista (MARCONI; LAKATOS, 2017).

Este estudo utiliza o método de pesquisa dedutivo, onde procura-se a todo custo confirmar a hipótese (GIL, 2008). As perguntas desenvolvidas e abordadas na introdução serão confirmadas ou falseadas de acordo com os resultados obtidos através das análises.

3.1.1 Objetivos da pesquisa

As pesquisas científicas podem ser classificadas em três tipos: exploratória, descritiva e explicativa, cada uma trata o problema de maneira diferente, dito isso, essa pesquisa pode ser classificada como exploratória, descritiva e explicativa.

Gil (2008) descreve que as pesquisas exploratórias têm como principal finalidade desenvolver, esclarecer e modificar conceitos e idéias, tendo em vista a formulação de problemas mais precisos ou hipóteses pesquisáveis para estudos posteriores. Envolvem levantamento bibliográfico e documental, entrevistas não padronizadas e estudos de caso, procedimentos de amostragem e técnicas quantitativas de coleta de dados não são costumeiramente aplicados nestas pesquisas.

A pesquisa descritiva, segundo Selltiz et al. (1965) ratificado por Oliveira (2011), busca descrever um fenômeno ou situação em detalhe, especialmente o que está ocorrendo, abrange com exatidão, as características de um indivíduo, uma situação, ou um grupo, bem como desvendar a relação entre os eventos. Uma de suas características mais significativas está na utilização de técnicas padronizadas de coleta de dados.

Gil (2008) define a pesquisa causal ou explicativa como baseada em experimentos, envolvendo hipóteses especulativas, definindo relações causais. A pesquisa explicativa tem como preocupação central identificar os fatores que determinam ou que contribuem para a ocorrência dos fenômenos. Este é o tipo de pesquisa que mais aprofunda o conhecimento da realidade, porque explica a razão, o porquê das coisas.

3.1.2 Natureza da pesquisa

A natureza dessa pesquisa é aplicada, tendo como objetivo gerar conhecimentos para aplicações práticas voltadas à solução de problemas específicos. Além disso, pode ser classificada também como qualitativa-quantitativa, pois além de reunir dados sobre alunos e ex-alunos da instituição, foram feitas interpretações e análises subjetivas, dando origem a um estudo completo sobre um objeto.

Almeida (2006) descreve a pesquisa qualitativa como uma relação existente entre o mundo e o sujeito que não pode ser traduzida em números, sendo classificada como descritiva, onde o pesquisador tende a analisar seus dados indutivamente. Já a pesquisa quantitativa considera que tudo é quantificável, o que significa traduzir opiniões e números em informações as quais serão classificadas e analisadas.

3.1.3 Objetivos de estudo

Considerando os objetivos desta pesquisa, pode-se classificá-la como explicativa, pois seu objetivo consta relacionar e explicar o desempenho dos alunos e ex-alunos com os indicadores de fluxo, identificando possíveis fatores determinantes ou contribuintes para o impacto no desempenho ao longo do curso.

3.2 População e amostra

A pesquisa tem como população os discentes do Instituto Federal de Educação Ciência e Tecnologia de Minas Gerais, e como amostra os alunos dos cursos superiores do *campus* São João Evangelista, referente ao período de 2017.1 a 2022.1 e que tenham respondido o questionário socioeconômico disponibilizado pelo instituto no início dos períodos letivos.

4. RESULTADOS E DISCUSSÃO

Neste capítulo, são apresentados os resultados e conclusões acerca das análises realizadas nas bases de dados do Instituto Federal de Educação, Ciências e Tecnologia de Minas Gerais campus São João Evangelista, utilizando os Indicadores de Fluxo como base para essas análises.

4.1 Contextualização do problema

O IFMG - SJE possui um currículo com 6 cursos de graduação nas modalidades de bacharelado e licenciatura, com um total de 334 disciplinas divididas ao longo dos semestres de acordo com seus respectivos cursos. O perfil dos ingressantes nos cursos são principalmente jovens adultos entre 18 a 24 anos, que ingressam na instituição através do vestibular ou pelo Sistema de Seleção Unificada (SISU).

Como fonte de dados para este trabalho, foram utilizados dados acadêmicos (BASE_ACADEMICO) e socioeconômicos (BASE_SOCIECONOMICO) dos alunos e ex-alunos dos cursos de Administração, Agronomia, Ciências Biológicas, Engenharia Florestal, Matemática e Sistemas de Informação do IFMG - SJE. Os dados fornecidos pelo setor de informática do *campus* não possuem forma de identificação do aluno, assim preservando os dados pessoais nas análises relacionadas aos indicadores de fluxo e definição de *personas*, de acordo com a Lei Geral de Proteção de Dados (LGPD).

A BASE_ACADEMICO foi disponibilizada em formato de arquivo Excel (xlsx), contendo as seguintes colunas:

i) COD_PESSOA (código da pessoa);

ii) DTNASCIMENTO (data de nascimento da pessoa);

iii) IDADE ATUAL (idade atual da pessoa);

iv) SEXO (sexo da pessoa);

v) MODALIDADE_CURSO (modalidade referente ao curso);

vi) NOME_CURSO (nome do curso);

vii) CODGRADE (código da grade referente ao curso);

viii) CODPERLET (código do período letivo);

ix) CODTURMA (código da turma);

x) NOME_DISCIPLINA (nome da disciplina);

xi) Nº REQUISITOS (número de requisitos solicitados por uma disciplina);

xii) NOTA_FINAL (nota final da disciplina);

xiii) TOTAL_FALTAS (total de faltas na disciplina por pessoa);

xiv) CARGA_HORARIA_DISCIPLINA (carga horária da disciplina);

xv) STATUS_DISCIPLINA (status da disciplina por pessoa);

xvi) STATUS_CURSO (status da pessoa referente ao curso).

A BASE_SOCIOECONOMICO também foi disponibilizada em formato de arquivo Excel (xlsx), contendo as seguintes colunas:

i) CODIGO (código da pessoa);

ii) CODQUESTAO (código referente a questão do questionário socioeconômico enumerado de 6 ao 15);

iii) QUESTAO (descrição da questão);

iv) RESPOSTA (resposta referente a questão).

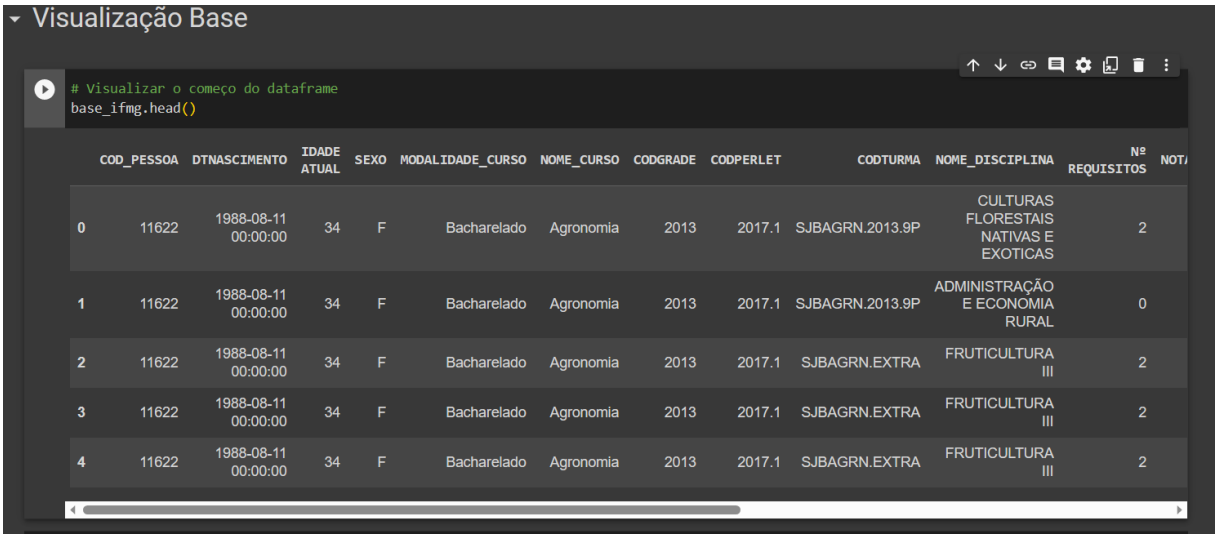
O primeiro passo do trabalho foi um estudo detalhado dos dados acadêmicos e socioeconômicos cedidos pelo setor de informática do IFMG - SJE, feito em duas etapas: Em um primeiro momento analisando diretamente os dados, após a importação dos mesmos no Google Collab, aplicando técnicas de seleção, limpeza e tratamento dos dados, com o objetivo de prepará-los para a etapa de mineração de dados e também possibilitando que os mesmos contemplassem as informações históricas referentes aos alunos em seus períodos no instituto.

Em um segundo momento, os dados foram submetidos à etapa de mineração de dados. A associação será considerada como tarefa base deste trabalho, juntamente com o agrupamento ou clusterização, também amplamente utilizados neste contexto.

4.2 Domínio da base de dados

A Figura 5 mostra a consulta retornando os dados da BASE_ACADEMICO e a Figura 6 os dados da BASE_SOCIOECONOMICO. Os demais arquivos texto e consultas podem ser vistos no APÊNDICE A.

Figura 5 - Visualização BASE_ACADEMICO



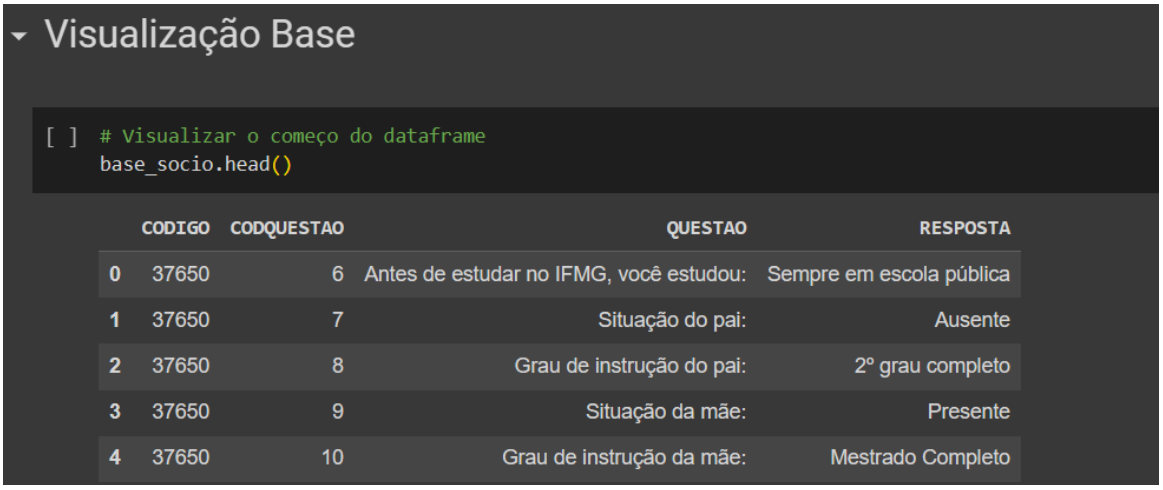
Visualização Base

```
# Visualizar o começo do dataframe
base_ifmg.head()
```

	COD_PESSOA	DTNASCIMENTO	IDADE ATUAL	SEXO	MODALIDADE_CURSO	NOME_CURSO	CODGRADE	CODPERLET	CODTURMA	NOME_DISCIPLINA	Nº REQUISITOS	NOTA
0	11622	1988-08-11 00:00:00	34	F	Bacharelado	Agronomia	2013	2017.1	SJBAGRN.2013.9P	CULTURAS FLORESTAIS NATIVAS E EXOTICAS	2	
1	11622	1988-08-11 00:00:00	34	F	Bacharelado	Agronomia	2013	2017.1	SJBAGRN.2013.9P	ADMINISTRAÇÃO E ECONOMIA RURAL	0	
2	11622	1988-08-11 00:00:00	34	F	Bacharelado	Agronomia	2013	2017.1	SJBAGRN.EXTRA	FRUTICULTURA III	2	
3	11622	1988-08-11 00:00:00	34	F	Bacharelado	Agronomia	2013	2017.1	SJBAGRN.EXTRA	FRUTICULTURA III	2	
4	11622	1988-08-11 00:00:00	34	F	Bacharelado	Agronomia	2013	2017.1	SJBAGRN.EXTRA	FRUTICULTURA III	2	

Fonte: Elaborado pelos autores, 2023.

Figura 6 - Visualização BASE_SOCIECONOMICO



Visualização Base

```
[ ] # Visualizar o começo do dataframe
    base_socio.head()
```

	CODIGO	CODQUESTAO	QUESTAO	RESPOSTA
0	37650	6	Antes de estudar no IFMG, você estudou:	Sempre em escola pública
1	37650	7	Situação do pai:	Ausente
2	37650	8	Grau de instrução do pai:	2º grau completo
3	37650	9	Situação da mãe:	Presente
4	37650	10	Grau de instrução da mãe:	Mestrado Completo

Fonte: Elaborado pelos autores, 2023.

O APÊNDICE B exibe os atributos das bases de dados que foram disponibilizados em formato de arquivo excel (xlsx) para o estudo e o domínio de valores aplicados a ele, caso exista. A BASE_ACADEMICO possui 16 colunas totalizando 38.707 registros e a BASE_SOCIECONOMICO possui 04 colunas totalizando 6.890 registros.

Seguindo as etapas do processo de KDD, o passo seguinte foi a limpeza e tratamento da BASE_ACADEMICO, dessa forma foi realizado a remoção de dados duplicados e o tratamento das inconsistências e dados faltantes, conforme descrito nas Figuras 7 e 8.

Figura 7 - Resultado de remoção dos dados duplicados

```
duplicados = base_ifmg[base_ifmg.duplicated(keep='first')]
print(duplicados)

Empty DataFrame
Columns: [COD_PESSOA, IDADE ATUAL, SEXO, MODALIDADE_CURSO, NOME_CURSO, CODPERLET, CODTURMA, NOME_DISCIPLINA, Nº REQUISIT
Index: []
```

Fonte: Elaborado pelos autores, 2023.

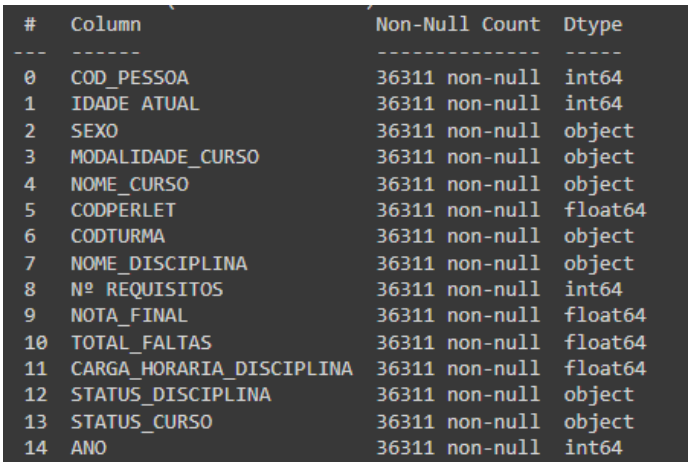
Figura 8 - Remoção das inconsistências e preenchimento dos dados faltantes

```
[673] i = base_ifmg.groupby(['COD_PESSOA', 'STATUS_DISCIPLINA']).count()
      agrupar_not_cod = i.groupby(['STATUS_DISCIPLINA']).size()
      agrupar_not_cod

      STATUS_DISCIPLINA
      Aprovado                  1264
      Desligado                    5
      Evasão                       5
      Exame Final                  1
      Formado                      1
      Matriculado (PD)           145
      Reprovado                  936
      Reprovado por frequência   764
      Trancado                     6
      dtype: int64
```

Fonte: Elaborado pelos autores, 2023.

Ainda nessa etapa, houve a exclusão dos atributos que não teriam influência sobre as análises, sendo as colunas de DTNASCIMENTO e CODGRADE, assim como as modalidades de pós-graduação e tecnologia, além disso foi criado a coluna ANO para auxiliar na etapa de mineração, conforme Figura 9.

Figura 9 - Visualização dos atributos na BASE_ACADEMICO pós processamento



#	Column	Non-Null Count	Dtype
0	COD_PESSOA	36311 non-null	int64
1	IDADE ATUAL	36311 non-null	int64
2	SEXO	36311 non-null	object
3	MODALIDADE_CURSO	36311 non-null	object
4	NOME_CURSO	36311 non-null	object
5	CODPERLET	36311 non-null	float64
6	CODTURMA	36311 non-null	object
7	NOME_DISCIPLINA	36311 non-null	object
8	Nº REQUISITOS	36311 non-null	int64
9	NOTA_FINAL	36311 non-null	float64
10	TOTAL_FALTAS	36311 non-null	float64
11	CARGA_HORARIA_DISCIPLINA	36311 non-null	float64
12	STATUS_DISCIPLINA	36311 non-null	object
13	STATUS_CURSO	36311 non-null	object
14	ANO	36311 non-null	int64

Fonte: Elaborado pelos autores, 2023.

Por fim, na etapa de transformação de dados a coluna CODPERLET onde encontram-se os períodos letivos, como por exemplo 2017.1, foi transformada em uma coluna denominada ANO para facilitar nas análises referentes aos cenários antes, durante e pós pandemia.

Vale destacar que esse processo não foi realizado na BASE_SOCIOECONOMICO, uma vez que as pequenas alterações foram realizadas no próprio arquivo excel (xlsx), não sendo necessário esforços em tratamento.

Após as etapas anteriores, realizou-se uma análise exploratória tanto dos dados categóricos, quanto dos dados numéricos. Os dados categóricos são valores para uma variável qualitativa, geralmente um número, uma palavra ou um símbolo, já os dados numéricos são basicamente os dados quantitativos obtidos de uma variável, e o valor tem uma sensação de tamanho / magnitude. (GOLDSCHMIDT; PASSOS, 2005).

Essa análise foi crucial para a etapa de mineração de dados, onde foi possível se ter uma visão ampla do conjunto de dados a ser analisado. No APÊNDICE C e APÊNDICE D estão disponibilizados os códigos utilizados para realizar a análise exploratória dos dados.

4.3 Etapa de mineração

Nesta seção, são apresentadas as tarefas de mineração identificadas para serem aplicadas nas bases de dados, a preparação desses dados para serem utilizados nas tarefas

escolhidas, os experimentos que foram realizados, bem como uma breve discussão sobre os resultados obtidos nos experimentos.

Antes de realizar a fase da mineração foram estabelecidos alguns critérios para a realização das análises. O primeiro fator foi a pandemia, sendo analisado e definido o perfil dos alunos no cenário antes, durante e pós-pandemia, o segundo critério foi realizar todas as análises com base nos indicadores de fluxo, sendo eles desistência, permanência e conclusão. Por fim, foram selecionados os atributos da BASE_ACADEMICO que contribuíssem para a identificação dos perfis, sendo: COD_PESSOA, STATUS_CURSO, IDADE ATUAL, NOME_DISCIPLINA, SEXO, TOTAL_FALTAS e STATUS_DISCIPLINA, assim como os atributos da BASE_SOCIOECONOMICO sendo CODIGO, CODQUESTAO, QUESTÃO e RESPOSTA.

4.3.1 Aplicação de mineração de dados nas bases

Para a identificação das tarefas aplicáveis às bases de dados, além de sua análise e entendimento, foi feito um estudo nos trabalhos relacionados. As principais tarefas definidas para o estudo são descritas a seguir:

Para a realização das análises utilizou-se:

a) Classificação por indicadores de fluxo (desistência, permanência e conclusão), no intuito de descobrir o perfil dos alunos (*personas*) de acordo com os cenários;

b) Associações com os cenários de antes, durante e pós pandemia, na tentativa de identificar o perfil dos alunos e se a pandemia foi um fator relevante de impacto no desempenho;

c) Classificação com base nos atributos de idade, quantidade média de disciplinas cursadas, sexo, média da frequência por disciplina utilizando o total de faltas e a quantidade média de reprovações por disciplina para identificar padrões que resultam o aluno a desistir, permanecer e/ou concluir o curso.

4.4 Experimentos

4.4.1 Experimento A

Para a realização do experimento A, foi criada uma *persona* para cada indicador de fluxo, contendo sexo, idade, quantidade de disciplinas cursadas, frequência nas disciplinas e quantidade de matérias que reprovaram, esses dados foram obtidos através de analise para cada perfil, assim como dados socioeconômicos, contendo tipo de instituição cursada antes de entrar no IFMG-SJE, situação dos pais, escolaridade dos pais, residência, área de procedência, renda familiar e número de pessoas que compõe a família. O APÊNDICE E mostra os códigos desenvolvidos para essas análises.

A *persona* A para o indicador de desistência tem 25 anos, do sexo masculino, ele cursou 17 disciplinas durante sua permanência na faculdade, tem em média 17 faltas por matéria e 09 reprovações. A Persona A sempre estudou em escola pública, tendo seus pais presentes e reside com os mesmos. O pai possui ensino superior completo e sua mãe 2° grau completo, a residência é alugada por eles em área urbana, a família é composta por 04 pessoas e possui renda média de 03 salários.

Já a *persona* B para o indicador de permanência tem 24 anos, do sexo feminino, ela cursou 42 disciplinas durante a faculdade, tem em média 06 faltas por matéria e 08 reprovações. A Persona B sempre estudou em escola pública, tendo seus pais presentes e reside com os mesmos. O pai e mãe possuem 1° grau incompleto, a residência é própria em área urbana, a família é composta por 04 pessoas e possui renda média de 02 salários.

Por fim a *persona* C para o indicador de conclusão tem 27 anos, do sexo feminino, tendo concluído todas as disciplinas do curso, com média de 04 faltas por matéria e 02 reprovações. Como o N amostral é pequeno, com apenas uma amostra, não foi possível realizar teste estatístico na base socioeconômica para a *persona* C.

4.4.2 Experimento B

Para a realização do experimento B, foi criada uma persona para os cenários antes, durante e pós-pandemia de acordo com os indicador de fluxo, contendo nome, sexo, idade, quantidade de disciplinas cursadas, frequência nas disciplinas e quantidade de matérias que reprovaram, nesse experimento não foi utilizado a BASE_SOCIECONOMICO devido falta de dados que correpondessem a BASE_ACADEMICO. O APÊNDICE F mostra os códigos desenvolvidos para chegar às informações que serão apresentadas abaixo.

4.4.2.1 Antes da pandemia (2017.1 - 2019.2)

A *persona* D para o indicador de desistência tem 26 anos, do sexo masculino, ele cursou 16 disciplinas durante sua permanência na faculdade, tem em média 16 faltas por matéria e 09 reprovações.

A *persona* E para o indicador de permanência tem 25 anos, do sexo masculino, ela cursou até então 28 disciplinas na faculdade, tem em média 10 faltas por matéria e 08 reprovações.

A *persona* F para o indicador de conclusão tem 27 anos, do sexo feminino, tendo concluído todas as disciplinas do curso, tem em média 04 faltas por matéria e 02 reprovações.

4.4.2.2 Durante a pandemia (2020.1 - 2021.2)

A *persona* G para o indicador de desistência tem 24 anos, do sexo masculino, ele cursou 12 disciplinas durante sua permanência na faculdade, tem em média 23 faltas por matéria e 08 reprovações.

A *persona* H para o indicador de permanência tem 24 anos, do sexo feminino, ele cursou até então 19 disciplinas na faculdade, tem em média 06 faltas por matéria e 05 reprovações.

A *persona* I para o indicador de conclusão tem 26 anos, do sexo masculino, tendo concluído todas as disciplinas do curso, tem em média 01 falta por matéria e nenhuma reprovação.

4.4.2.3 Pós-pandemia (2022.1)

A *persona* J para o indicador de desistência tem 24 anos, do sexo masculino, ele cursou 04 disciplinas durante sua permanência na faculdade, tem em média 12 faltas por matéria e 04 reprovações.

A *persona* K para o indicador de permanência tem 23 anos, do sexo feminino, ela cursou até então 05 disciplinas na faculdade, tem em média 08 faltas por matéria e 02 reprovações.

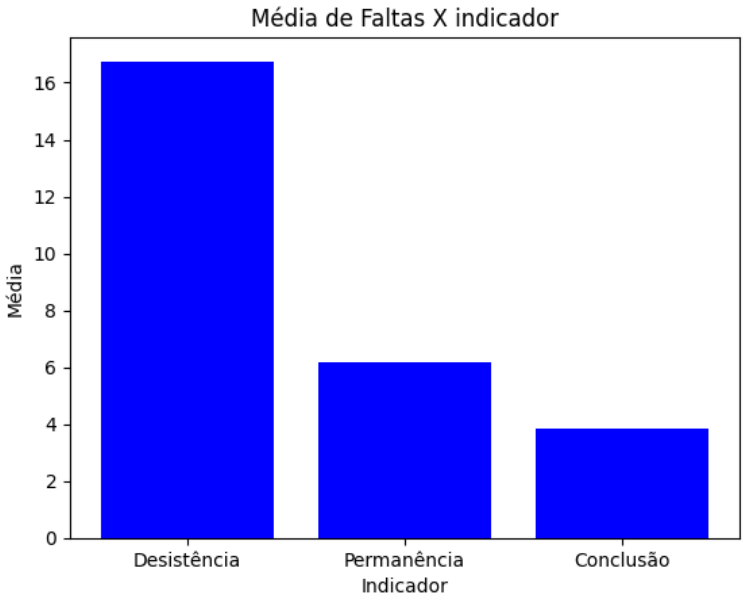
A *persona* L para o indicador de conclusão tem 24 anos, do sexo masculino, tendo concluído todas as disciplinas do curso, tem em média 04 faltas por matéria e nenhuma reprovação.

4.5 Discussão dos resultados obtidos

Para o indicador de desistência, foram considerados os alunos que tenham o status do curso como: Trancado, Evasão ou Desligado. Os gráficos relacionados ao experimento A e B podem ser visualizados nos APÊNDICE G e H.

O objetivo do experimento A foi investigar o perfil geral dos alunos de acordo com cada indicador de fluxo, levando em consideração todos os registros disponíveis na base de dados. De acordo com os resultados obtidos, um dos fatores que chama a atenção é a quantidade média de faltas que gira em torno de 17 para quem tem tendência a desistir do curso, enquanto para os perfis de permanência e conclusão, totalizam em média 06 e 07, respectivamente.

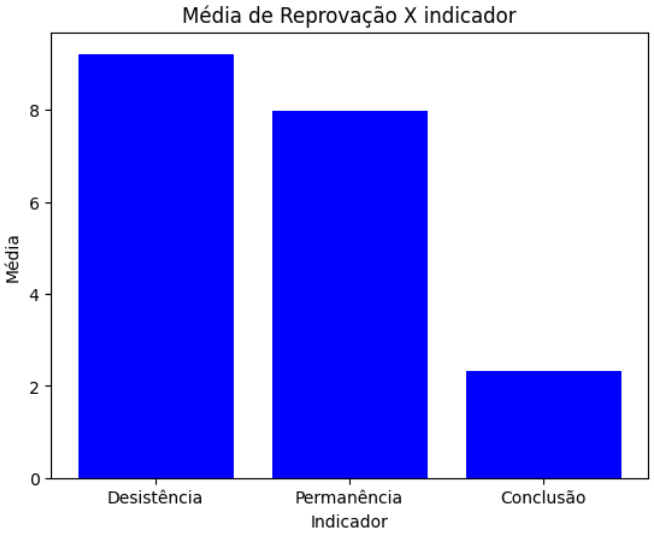
Figura 10 - Média de faltas por disciplinas e indicador



Fonte: Elaborado pelos autores, 2023.

Outro fator de atenção é o número de reprovações, em média as pessoas com tendência para desistência somam 09 reprovações, tendo cursado poucas disciplinas durante sua permanência no curso, já o perfil de permanência soma em média 08 reprovações, tendo cursado em média 42 disciplinas e para o perfil de conclusão temos em média 02 reprovações. Em relação à questão socioeconômica dos alunos, de acordo com os dados disponibilizados na BASE_SOCIOCONOMICO não foi possível identificar impacto, uma vez que os resultados se assemelham em todos os indicadores.

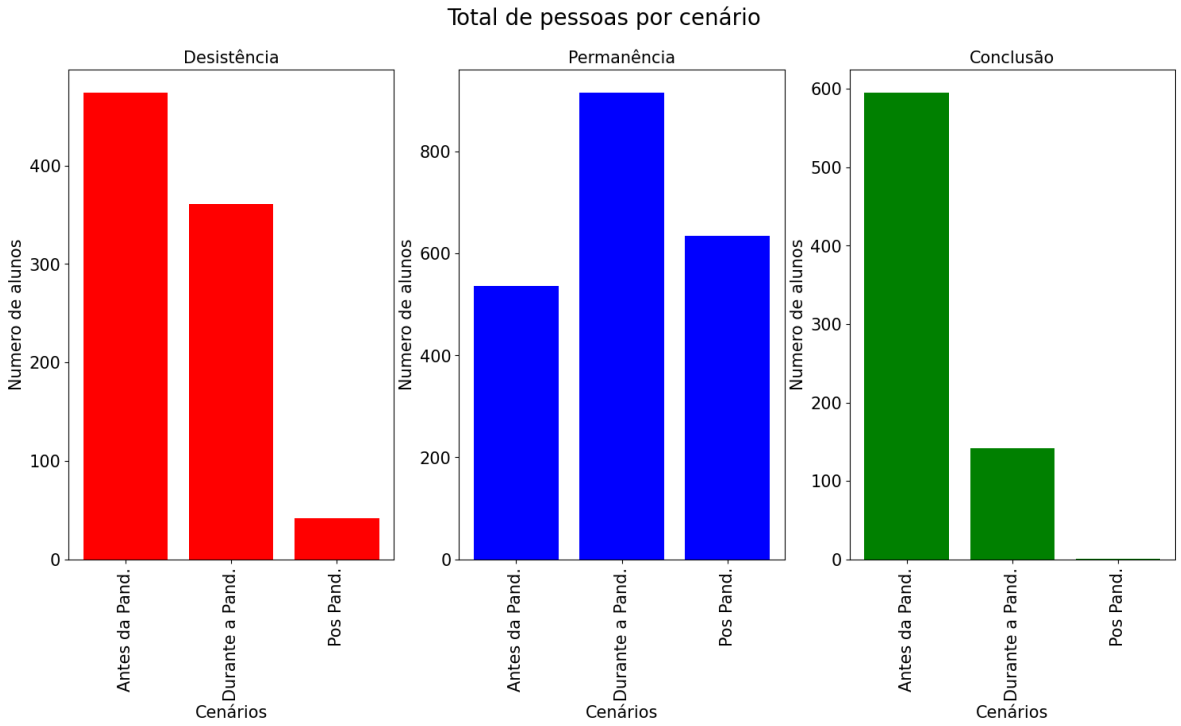
Figura 11 - Média de reprovações por disciplinas e indicador



Fonte: Elaborado pelos autores, 2023.

Para o experimento B, foram considerados além dos indicadores de fluxo, os cenários antes, durante e pós-pandemia. De maneira geral os resultados se assemelham e fortalecem os resultados do experimento A, no entanto de acordo com as análises realizadas o cenário de pandemia não se comportou de maneira exponencialmente diferente do esperado para cada indicador.

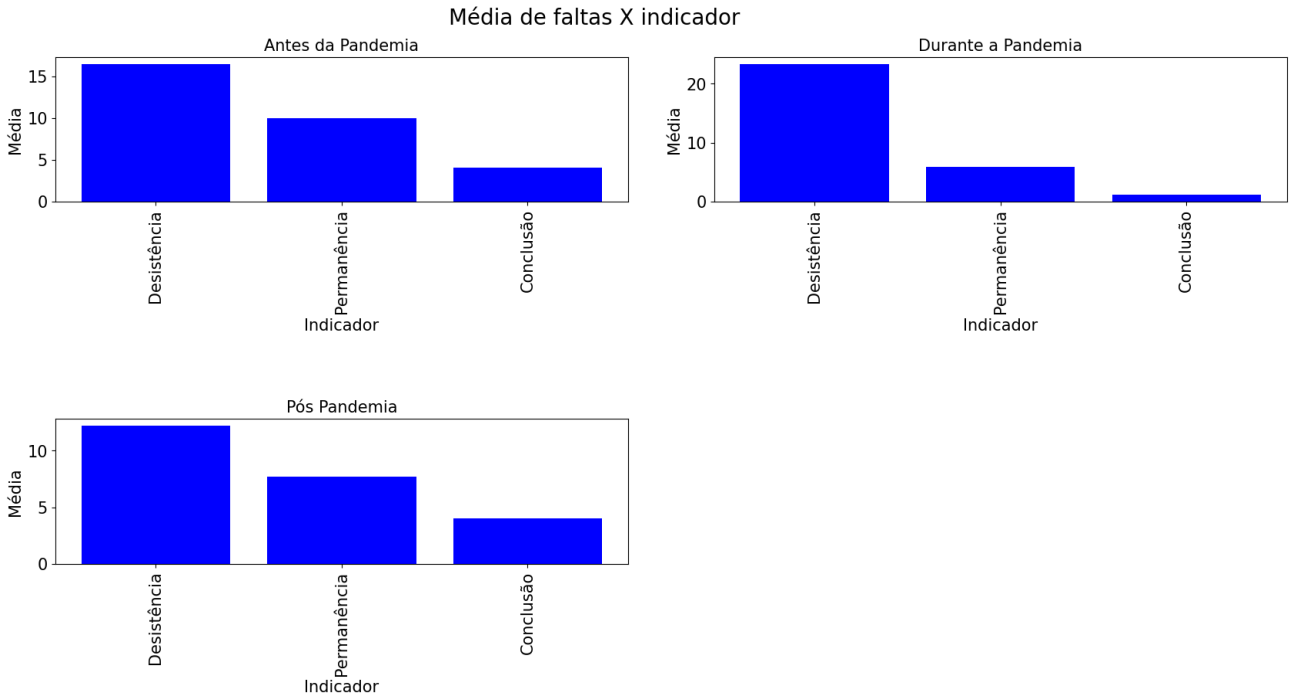
Figura 12 - Total de pessoas por cenário e indicadores.



Fonte: Elaborado pelos autores, 2023.

O principal destaque do experimento B foi o aumento da quantidade de faltas durante a pandemia para o perfil de desistência, totalizando em média 23 faltas. Para esse experimento não foram utilizados os dados socioeconômicos por serem considerados insuficientes.

Figura 13 - Média de reprovações por disciplinas, indicador e cenário



Fonte: Elaborado pelos autores, 2023.

Portanto, após análise detalhada de cada perfil definido, a resposta para as seguintes perguntas foram:

a) A quantidade de reprovação nas disciplinas impacta os indicadores de fluxo? Ficou evidente que o número de reprovações é um padrão para desistência, quanto mais reprovações ao longo do curso o aluno fica mais propício a desistir, já pessoas que reprovam poucas vezes tem mais chance de permanecer e concluir o curso.

b) A quantidade de faltas nas disciplinas impacta os indicadores de fluxo? Sim, pode-se perceber que pessoas que acumulam um grande número de faltas estão mais propícias a desistir do curso, enquanto pessoas com uma quantidade menor permanecem ou concluem o curso na maioria dos casos.

c) Qual gênero é mais propenso a desistir do curso? De acordo com as análises, pessoas do sexo masculino são mais propensas a desistirem do curso, enquanto as mulheres na maioria das vezes permanecem ou concluem o curso.

d) A pandemia de alguma forma afetou os indicadores de fluxo? Não foi possível identificar grande variações nos cenários, o comportamento antes, durante e pós-pandemia se assemelha ao comportamento do experimento A, com destaque para o aumento das faltas no

cenário durante a pandemia associado ao indicador de desistência e algumas mudanças no sexo do perfis durante os cenários.

e) Os dados socioeconômicos influenciam nos indicadores? De acordo com os levantamentos realizados o fator socioeconômico não influencia nos padrões de desempenho ao longo do curso, os resultados se assemelham em cada indicador.

5. CONSIDERAÇÕES FINAIS

Ao longo do desenvolvimento deste trabalho, foram realizados estudos em projetos semelhantes a este, visando atingir os objetivos propostos e atender os requisitos. Com base nas análises realizadas foi possível identificar padrões que possivelmente contribuem para o aluno evadir, permanecer e concluir o curso, com esses dados é possível que o instituto desenvolva ações direcionadas para minimizar ou evitar que o aluno abandone os estudos.

Algumas decisões e situações no decorrer do desenvolvimento do trabalho acarretaram retrabalho, perda de tempo, e dificuldade no andamento do estudo. Outras auxiliaram na constatação da relação entre as necessidades de conhecimentos teóricos e as dificuldades da realização na prática, como por exemplo, a falta de conhecimento sobre as bases, a criação de filtros para os dados, montagem dos gráficos para facilitar as análises, que tomaram mais tempo do que o esperado. Recomenda-se que para projetos futuros seja feita uma análise prévia do universo de dados que irá ser utilizado, antes de, por exemplo, à montagem de um cronograma. Também deve ser considerado um estudo mais profundo das bases antes do início do desenvolvimento do projeto.

Em uma primeira análise, pensou ser possível categorizar por período letivo ou turma (CODPERLET e CODTURMA), porém não é possível analisar o período letivo, uma vez que o único período letivo que está disponível é o 2022.1. Já o CODTURMA, poderia ser obtido o curso, ano e período letivo, se fosse analisado, porém poderia retornar um falso resultado, pois quando há uma situação em que um aluno de um determinado curso reprova em uma matéria, poderá se matricular na mesma matéria, em outro curso. Com isso a coluna CODTURMA assumiria o curso que aquela matéria está sendo disponibilizada.

Os dados socioeconômicos não são obrigatórios para responder, além disso, o preenchimento era realizado em papel impresso e recentemente foi substituído pelo formato digital, dessa forma as análises podem não ser tão precisas devido a quantidade de dados existentes para realizar a associação entre as bases.

Por fim, conclui-se que todos os objetivos foram alcançados, onde foi possível fazer as análises propostas e obter informações relevantes. Também foi possível aprimorar os conhecimentos em análise de dados, *data mining,* por meio do desenvolvimento desse estudo.

Para futuros estudos deixamos algumas propostas, como:

a) Análise mais profunda e com a base socioeconômica mais completa, para verificar a fundo o impacto socioeconômico na vida do aluno.

b) Analisar desde à entrada do aluno, à sua jornada até a conclusão do curso, e as principais matérias que impactam na sua permanência no curso.
c) Explorar se reprovação por frequência resulta em evasão, se o total de faltas influencia na nota do aluno, as diferenças das taxas de aprovação, e reprovação entre as modalidades (bacharelado e licenciatura), e à identificação do perfil por curso com maior chance de não conclusão.
d) Criação de um banco de dados para armazenar os dados de forma que fique fácil analisar, e que os algoritmos de mineração exigem, possibilitando uma melhor visualização dos resultados;
e) Criação de uma ferramenta para os coordenadores utilizarem, onde possam fazer a mineração dos dados gerados ao longo do curso pelos alunos, facilitando a gestão dos mesmos.

REFERÊNCIAS

ABMES - Revista Da Associação Brasileira De Mantenedoras De Ensino Superior (Brasil). **Indicadores de qualidade da Educação Superior. Aplicabilidade nas modalidades presencial e a distância**, [*s. l.*], ano 31, n. 43, Junho 2019. Disponível em: https://abmes.org.br/arquivos/publicacoes/miolo_estudos43_13052019.pdf. Acesso em: 20 maio 2022.

ALMEIDA, Maurício B. **Noções básicas sobre metodologia de pesquisa científica.** DTGI ECI/UFMG, 2006. Disponível em: https://mba.eci.ufmg.br/downloads/metodologia.pdf. Acesso em: 30 maio 2022.

BAKER, Ryan Shaun Joazeiro de; ISOTANI, Seiji; CARVALHO, Adriana Maria Joazeiro Baker de. **Mineração de Dados Educacionais: Oportunidades para o Brasil.** Revista Brasileira de Informática na Educação, [*s. l.*], v. 19, n. 2, 2011. Disponível em: http://ojs.sector3.com.br/index.php/rbie/article/view/1301/1172. Acesso em: 21 maio 2022.

BERNARDINI, Flavia Cristina. **Introdução ao Processo de KDD e MD.** Disponível em: https://www.professores.uff.br/fcbernardini/wp-content/uploads/sites/68/2017/08/01-Introdu%C3%A7%C3%A3o-a-KDD-e-DM.pdf . Acesso em 01 jun. 2022.

BORGES, Luis Eduardo. **Python para desenvolvedores.** 2ª edição. Disponível em: <https://ark4n.files.wordpress.com/2010/01/python_para_desenvolvedores_2ed.pdf>. Acesso em: 23 de abr. de 2022. Rio de Janeiro. Edição do Autor, 2010.

BRASIL. **Lei Nº 9.394, de 20 de Dezembro de 1996.** Estabelece as diretrizes e bases da educação nacional. Brasília: Presidência da República, [1996]. Disponível em: http://www.planalto.gov.br/ccivil_03/leis/l9394.htm. Acesso em: 01 jun. 2022.

BRASIL. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). **Censo da Educação Superior 2020: notas estatísticas.** Brasília, DF: Inep, 2022. Disponível em:

https://download.inep.gov.br/publicacoes/institucionais/estatisticas_e_indicadores/notas_estatisticas_censo_da_educacao_superior_2020.pdf. Acesso em: 20 maio 2020.

CAMPOS, Roger Júnio. **Previsão de séries temporais com aplicações a séries de consumo de energia elétrica.** 2008. Dissertação (Pós Graduação em Engenharia Elétrica) - Universidade Federal de Minas Gerais, Belo Horizonte, 2008. Disponível em: https://repositorio.ufmg.br/bitstream/1843/BUOS-8CTETD/1/290m.pdf. Acesso em: 25 maio 2022.

CORNELIUS JUNIOR, Romeu. **Uso da mineração de dados na identificação de alunos com perfil de evasão do ensino superior.** 2015. Disponível em: https://repositorio.unisc.br/jspui/bitstream/11624/535/1/Romeu%20Cornelius%20Junior%20-%20TCC%20-%20Final.pdf. Acesso em: 18 maio 2022.

CUNHA, Simone Miguez; CARRILHO, Denise Madruga. O processo de adaptação ao ensino superior e o rendimento acadêmico. **Psicologia escolar e educacional**, v. 9, p. 215-224, 2005. Disponível em: https://www.scielo.br/j/pee/a/qjznyDrBP5CtCf5MmLxZLgv/?lang=pt. Acesso em: 18 maio 2022.

DA COSTA CÔRTES, Sérgio; PORCARO, Rosa Maria; LIFSCHITZ, Sérgio. **Mineração de dados - funcionalidades, técnicas e abordagens.** [*S. l.*]: PUC, 2002. Disponível em: https://www.dbd.puc-rio.br/depto_informatica/02_10_cortes.pdf. Acesso em: 25 maio 2022.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From data mining to knowledge discovery in databases. **AI magazine**, v. 17, n. 3, p. 37-37, 1996. Disponível em: https://ojs.aaai.org/index.php/aimagazine/article/view/1230/1131. Acesso em: 20 maio 2022.

FONSACA, H. H.; DOROCINSKI, M. A. S. DEVIAS: **Mineração de dados educacionais para o auxílio à tomada de decisão.** 2018. Trabalho de conclusão de curso (Tecnólogo em Análise e Desenvolvimento de Sistemas) - Universidade Federal Do Paraná, [*S. l.*], 2018. Disponível em:

https://acervodigital.ufpr.br/bitstream/handle/1884/56646/Devias%20-%20Monografia.pdf?sequence=1&isAllowed=y. Acesso em: 20 maio 2022.

FURLAN, M. B. **Algoritmos e técnicas para Mineração de Dados.** 2018. Trabalho de conclusão de curso (Bacharelado em Ciência da Computação) - Fundação Educacional do Município de Assis, Assis, 2018. Disponível em: https://cepein.femanet.com.br/BDigital/arqTccs/1511420203.pdf. Acesso em: 18 maio 2022.

GIL, Antonio Carlos. **Métodos e Técnicas de Pesquisa Social**. 6. ed. São Paulo: Atlas S.A., 2008. ISBN 978-85-224-5142-5. *E-book*.

GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel. **Data Mining: um guia prático: Conceitos, técnicas, ferramentas, orientações e aplicações.** 4. ed. Rio de Janeiro: Elsevier, 2005. 262 p. ISBN 85-352-1877-7. *E-book*.

HAN, Jiawei; KAMBER, Micheline. **Data Mining: Concepts and Techniques.** 2. ed. [*S. l.*: *s. n.*], 2006. ISBN 978-1-55860-901-3. *E-book*.

INSTITUTO SEMESP (SP). **Dados Brasil.** 11. ed. [*S. l.*], 2021. Disponível em: https://www.semesp.org.br/mapa-do-ensino-superior/edicao-11/dados-brasil/evasao/. Acesso em: 20 maio 2022.

INSTITUTO SEMESP (Brasil). **Evasão bate recordes no ensino superior. Brasil: Desafios da Educação**, 25 jan. 2022. Disponível em: https://www.semesp.org.br/imprensa/evasao-bate-recordes-no-ensino-superior/. Acesso em: 25 maio 2022.

LIMA, Lucas Guilherme Pontes. **Mineração de dados utilizando a base de dados dos atendimentos do sistema nacional de emprego-SINE de Palmas/TO**. 2019. Disponível em: https://repositorio.uft.edu.br/bitstream/11612/2859/1/Lucas%20Guilherme%20Pontes%20Lima-%20TCC.pdf. Acesso em: 20 maio 2022.

MARCONI, Marina de Andrade; LAKATOS, Eva Maria. **Fundamentos de Metodologia Científica**. 8. ed. São Paulo: Atlas S.A., 2017. ISBN 978-85-970-1076-3. *E-book*.

MELO, G. S.; SALVADOR, T. L.; DE SOUZA, V. A. **Análise do comportamento do sistema de abastecimento e distribuição de água da cidade de Guanhães utilizando técnicas de Mineração de Dados**. 2014. Trabalho de conclusão de curso (Bacharelado em Sistemas da Informação) - Instituto Federal de Minas Gerais, [*S. l.*], 2014. Disponível em: https://www.sje.ifmg.edu.br/portal/images/artigos/biblioteca/TCCs/Sistemas_de_informacao/2014/GUSTAVO_SALVADOR_MELO_THIAGO_LEITE_SALVADOR_VICTOR_ALBERTO_DE_SOUZA.pdf. Acesso em: 18 maio 2022.

MINISTÉRIO DA EDUCAÇÃO (Brasil). DEED - Diretoria De Estatísticas Educacionais. **Metodologia de Cálculo dos Indicadores de Fluxo da Educação Superior**, [*s. l.*], 2017. Disponível em: https://download.inep.gov.br/informacoes_estatisticas/indicadores_educacionais/2017/metodologia_indicadores_trajetoria_curso.pdf. Acesso em: 20 maio 2022.

MINISTÉRIO DA EDUCAÇÃO (Brasil). INEP - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. **Indicadores de Fluxo da Educação Superior.** [*S. l.*], 21 out. 2020a. Disponível em: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/indicadores-educacionais/indicadores-de-fluxo-da-educacao-superior. Acesso em: 20 maio 2022.

MINISTÉRIO DA EDUCAÇÃO (Brasil). INEP - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. **Indicadores de Qualidade da Educação Superior.** [*S. l.*]. Disponível em: https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/indicadores-de-qualidade-da-educacao-superior. Acesso em: 25 maio 2022.

MINISTÉRIO DA EDUCAÇÃO (Brasil). INEP - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. **Censo da Educação Superior**. [*S. l.*]. Disponível em: https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/censo-da-educacao-superior. Acesso em: 25 maio 2022

OLIVEIRA, F. E. de. **Busca de conhecimento sobre o processo seletivo do Instituto Federal de Minas Gerais – Campus São João Evangelista: A mineração de dados como tecnologia para desvelar tendências e padrões.** 2013. Trabalho de conclusão de curso (Bacharelado em Sistemas da Informação) - Instituto Federal de Minas Gerais, São João Evangelista, 2013. Disponível em: https://www.sje.ifmg.edu.br/portal/images/artigos/biblioteca/TCCs/Sistemas_de_informacao/2013/FERNANDO_ELIAS_DE_OLIVEIRA.pdf. Acesso em: 20 maio 2022.

OLIVEIRA, M. F. de. (2011). **Metodologia científica: um manual para a realização de pesquisas em Administração**. Universidade Federal de Goiás. Catalão–GO. Disponível em: https://files.cercomp.ufg.br/weby/up/567/o/Manual_de_metodologia_cientifica_-_Prof_Maxwell.pdf. Acesso em: 02 jun. 2022

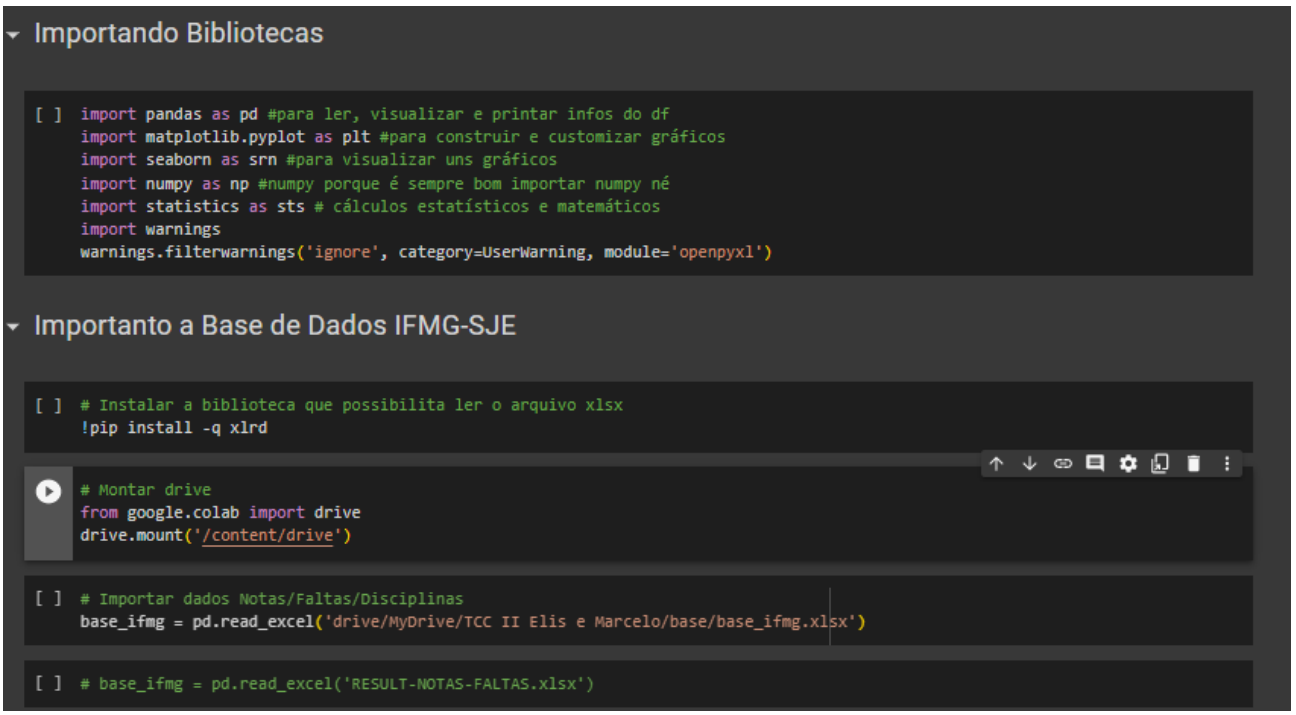
SANTOS, C. D. C. M. O. dos. **Análise de dados para avaliação do desempenho dos alunos do Instituto de Computação da UFF**. 2018. Disponível em: https://app.uff.br/riuff/bitstream/handle/1/5738/Monografia%20-%20Carlos%20Daniel.pdf;jsessionid=34C3427A536251892F00A8DFCCFB4235?sequence=1. Acesso em: 9 maio 2022.

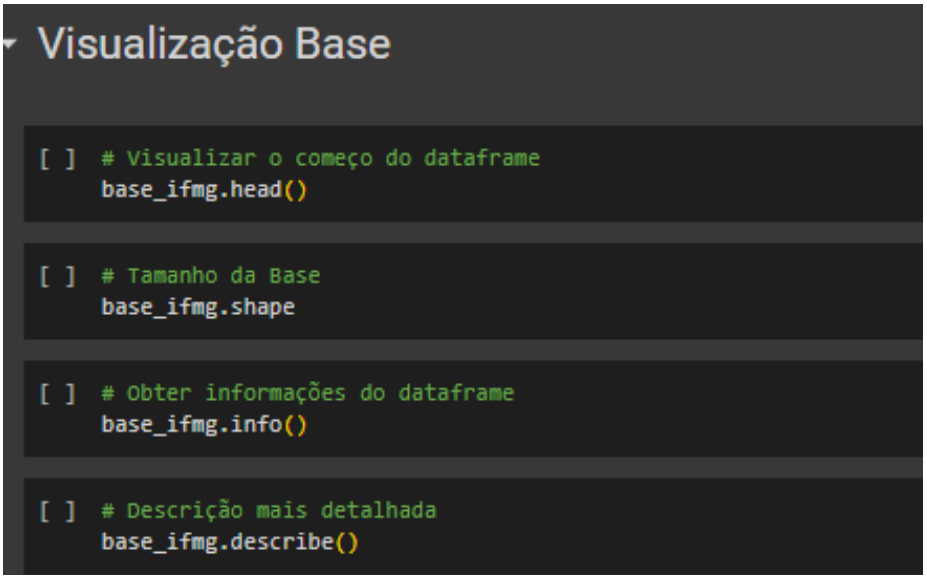
SILVA, Leandro Nunes de Castro; FERRARI, Daniel Gomes. **Introdução a Mineração de Dados: Conceitos básicos, algoritmos e aplicações**. São Paulo: Saraiva Educação, 2016. 673 p. ISBN 978-85-472-0099-2. *E-book*.

PATRÍCIO JÚNIOR, José Carlos Almeida. Mining Knowledge TV: **Uma abordagem de ambiente de KDD com ênfase em Mineração de Dados no ambiente da Knowledge TV. 2012. Dissertação (Pós Graduação em Informática) - Universidade Federal da Paraíba, João Pessoa**, 2012. Disponível em: https://repositorio.ufpb.br/jspui/bitstream/tede/6068/1/arquivototal.pdf. Acesso em: 25 maio 2022.

WITTEN, Ian H.; FRANK, Eibe. **Data Mining: Pratical Machine Learning Tools and Techniques**. 2. ed. [*S. l.*]: Elsevier Inc., 2005. ISBN 0-12-088407-0. *E-book*.

APÊNDICE A - Importando as bibliotecas, o arquivo de base, e comando para visualizar a BASE_ACADEMICO.





Fonte: Elaborado pelos autores, 2023.

APÊNDICE B - Atributos das bases de dados

```
base_ifmg.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 36311 entries, 0 to 38706
Data columns (total 16 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   COD_PESSOA               36311 non-null  int64
 1   IDADE ATUAL              36311 non-null  int64
 2   SEXO                     36311 non-null  object
 3   MODALIDADE_CURSO         36311 non-null  object
 4   NOME_CURSO               36311 non-null  object
 5   CODPERLET                36311 non-null  float64
 6   CODTURMA                 36311 non-null  object
 7   NOME_DISCIPLINA          36311 non-null  object
 8   Nº REQUISITOS            36311 non-null  int64
 9   NOTA_FINAL               36311 non-null  float64
 10  TOTAL_FALTAS             36311 non-null  float64
 11  CARGA_HORARIA_DISCIPLINA 36311 non-null  float64
 12  STATUS_DISCIPLINA        36311 non-null  object
 13  STATUS_CURSO             36311 non-null  object
 14  ANO                      36311 non-null  int64
 15  ESTA                     36311 non-null  object
dtypes: float64(4), int64(4), object(8)
memory usage: 4.7+ MB
```

```
[ ]  # Obter informações do dataframe
     base_socio.info()

     <class 'pandas.core.frame.DataFrame'>
     RangeIndex: 6890 entries, 0 to 6889
     Data columns (total 4 columns):
      #   Column      Non-Null Count  Dtype
     ---  ------      --------------  -----
      0   CODIGO      6890 non-null   int64
      1   CODQUESTAO  6890 non-null   int64
      2   QUESTAO     6890 non-null   object
      3   RESPOSTA    6890 non-null   object
     dtypes: int64(2), object(2)
     memory usage: 215.4+ KB
```

Fonte: Elaborado pelos autores, 2023.

APÊNDICE C - Códigos utilizados para realizar a análise exploratória na BASE_ACADEMICO

Sexo

```
[ ]  # Qtd de pessoas por sexo
     s = base_ifmg.groupby(['COD_PESSOA', 'SEXO']).count()
     agrupar_sexo = s.groupby(['SEXO']).size()
     agrupar_sexo
```

```
     agrupar_sexo.plot.bar(color = 'purple')
```

Modalidade cursos

```
     # Qtd de pessoas por modalidade curso
     m = base_ifmg.groupby(['COD_PESSOA', 'MODALIDADE_CURSO']).count()
     agrupar_modalidade = m.groupby(['MODALIDADE_CURSO']).size()
     agrupar_modalidade
```

```
     agrupar_modalidade.plot.bar(color = 'blue')
```

Cursos

```
[ ]  # Qtd de pessoas por curso
     c = base_ifmg.groupby(['COD_PESSOA', 'NOME_CURSO']).count()
     agrupar_curso = c.groupby(['NOME_CURSO']).size()
     agrupar_curso
```

```
[ ]  agrupar_curso.plot.bar(color = 'blue')
```

Turma

```
[ ]  # Qtd de pessoas por turma
     t = base_ifmg.groupby(['COD_PESSOA', 'CODTURMA']).count()
     agrupar_turma = t.groupby(['CODTURMA']).size()
     agrupar_turma
```

Disciplinas

```
[ ]  # Qtd de pessoas por disciplina
     d = base_ifmg.groupby(['COD_PESSOA', 'NOME_DISCIPLINA']).count()
     agrupar_disc = d.groupby(['NOME_DISCIPLINA']).size()
     agrupar_disc
```

```
[ ]  # Disciplinas cursadas por pessoa
     q = base_ifmg.groupby(['COD_PESSOA', 'NOME_DISCIPLINA']).count()
     agrupar_disc = q.groupby(['COD_PESSOA', 'NOME_DISCIPLINA']).size()
     agrupar_disc
```

Status disciplinas

```
# Qtd de pessoas por status da disciplinas
sd = base_ifmg.groupby(['COD_PESSOA', 'STATUS_DISCIPLINA']).count()
agrupar_sd = sd.groupby(['STATUS_DISCIPLINA']).size()
agrupar_sd
```

```
agrupar_sd.plot.bar(color = 'green')
```

Pessoas x Diciplinas x Status das disciplinas

```
[ ]  # Descrição pessoa, nome da disciplina e status da disciplina
     sd = base_ifmg.groupby(['COD_PESSOA', 'NOME_DISCIPLINA', 'STATUS_DISCIPLINA']).cc
     agrupar_sd = sd.groupby(['COD_PESSOA', 'NOME_DISCIPLINA', 'STATUS_DISCIPLINA']).s
     agrupar_sd
```

Status curso

```
[ ]  # Qtd de pessoas por status do curso
     sc = base_ifmg.groupby(['COD_PESSOA', 'STATUS_CURSO']).count()
     agrupar_sc = sc.groupby(['STATUS_CURSO']).size()
     agrupar_sc
```

```
agrupar_sc.plot.bar(color = 'red')
```

```
[ ]  # Descrição pessoa e status do curso
     sc = base_ifmg.groupby(['COD_PESSOA', 'STATUS_CURSO']).count()
     agrupar_sc = sc.groupby(['COD_PESSOA', 'STATUS_CURSO']).size()
     agrupar_sc
```

Requisitos disciplinas

```
[ ]  # Número de requisitos por disciplina
     r = base_ifmg.groupby(['Nº REQUISITOS', 'NOME_DISCIPLINA']).count()
     agrupar_req = r.groupby(['NOME_DISCIPLINA']).size()
     agrupar_req
```

```
[ ]  base_ifmg['Nº REQUISITOS'].describe()
```

Idade

```
[ ]  # Qtd de pessoas por idade
     i = base_ifmg.groupby(['COD_PESSOA', 'IDADE ATUAL']).count()
     agrupar_idade = i.groupby(['IDADE ATUAL']).size()
     agrupar_idade
```

```
     base_ifmg['IDADE ATUAL'].describe()
```

```
[ ]  srn.boxplot(x = base_ifmg['IDADE ATUAL']).set_title('IDADE ATUAL')
```

Período letivo

```
[ ]  # Qtd de pessoas por período letivo
     p = base_ifmg.groupby(['COD_PESSOA', 'CODPERLET']).count()
     agrupar_pl = p.groupby(['CODPERLET']).size()
     agrupar_pl
```

```
[ ]  srn.boxplot(x = base_ifmg['Nº REQUISITOS']).set_title('Nº REQUISITOS')
```

Disciplinas x Requisitos

```
[ ]  # Qtd de disciplina por requisito
     agrupar_reque = r.groupby(['Nº REQUISITOS']).size()
     agrupar_reque
```

```
[ ]  agrupar_reque.plot.bar(color = 'blue')
```

Nota final x Faltas por disciplina x Pessoas

```
# Nota final e total de faltas por disciplina e pessoa
n = base_ifmg.groupby(['COD_PESSOA', 'NOME_DISCIPLINA', 'NOTA_FINAL', 'TOTAL_FALT
agrupar_req = base_ifmg.groupby(['COD_PESSOA', 'NOME_DISCIPLINA', 'NOTA_FINAL', '
agrupar_req
```

Carga Horária

```
# Carga horária da disciplina
agrupar_carga = base_ifmg.groupby(['CARGA_HORARIA_DISCIPLINA']).size()
agrupar_carga
```

```
agrupar_carga.plot.bar(color = 'gray')
```

Faltas

```
# Total de Faltas
agrupar_faltas = base_ifmg.groupby(['TOTAL_FALTAS']).size()
agrupar_faltas
```

```
base_ifmg['TOTAL_FALTAS'].describe()
```

```
srn.boxplot(x = base_ifmg['TOTAL_FALTAS']).set_title('TOTAL FALTAS')
```

```
ax = srn.displot(base_ifmg.TOTAL_FALTAS, kde = False)
ax.figure.set_size_inches(12,6)
ax.set(title ="TOTAL FALTAS", xlabel='FALTAS')
ax
```

Administração

```
# Média do curso administração
b=base_ifmg[base_ifmg['NOME_CURSO'] == 'Administração']
nf = b['NOTA_FINAL'].mean()
nf
```

```
# Média de nota por curso e período letivo de administração
b = base_ifmg[base_ifmg['NOME_CURSO'] == 'Administração']
c = b.groupby(['CODTURMA', 'NOTA_FINAL'], as_index=False).mean().groupby('CODTURM
c
```

```
c.plot.bar(color = 'blue', figsize= (20, 10))
```

Sistemas de Informação

```
[ ]  # Média do curso sistemas de informação
     b=base_ifmg[base_ifmg['NOME_CURSO'] == 'Sistemas de Informação']
     nf = b['NOTA_FINAL'].mean()
     nf
```

```
[ ]  # Média de nota por curso e período letivo de sistemas de informação
     b = base_ifmg[base_ifmg['NOME_CURSO'] == 'Sistemas de Informação']
     c = b.groupby(['CODTURMA', 'NOTA_FINAL'], as_index=False).mean().groupby('CODTURM
     c
```

```
[ ]  c.plot.bar(color = 'blue', figsize= (20, 10))
```

Ciências Biológicas

```
[ ]  # Média do curso ciências biológica
     b=base_ifmg[base_ifmg['NOME_CURSO'] == 'Ciências Biológicas']
     nf = b['NOTA_FINAL'].mean()
     nf
```

```
[ ]  # Média de nota por curso e período letivo de ciências biológicas
     b = base_ifmg[base_ifmg['NOME_CURSO'] == 'Ciências Biológicas']
     c = b.groupby(['CODTURMA', 'NOTA_FINAL'], as_index=False).mean().groupby('CODTURM
     c
```

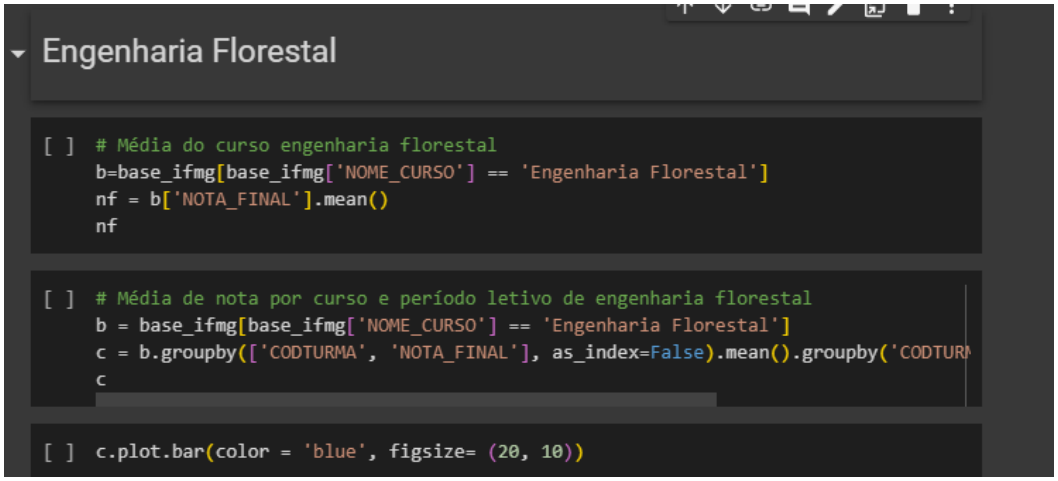
```
[ ]  c.plot.bar(color = 'blue', figsize= (20, 10))
```

Matemática

```
[ ]  # Média do curso matemática
     b=base_ifmg[base_ifmg['NOME_CURSO'] == 'Matemática']
     nf = b['NOTA_FINAL'].mean()
     nf
```

```
[ ]  # Média de nota por curso e período letivo de matemática
     b = base_ifmg[base_ifmg['NOME_CURSO'] == 'Matemática']
     c = b.groupby(['CODTURMA', 'NOTA_FINAL'], as_index=False).mean().groupby('CODTURM
     c
```

```
[ ]  c.plot.bar(color = 'blue', figsize= (20, 10))
```



Fonte: Elaborado pelos autores, 2023.

APÊNDICE D - Códigos utilizados para realizar a análise exploratória na BASE_SOCIOECONOMICO.

Questão 6

```
[ ]  # Qtd de pessoas por resposta
     print("Antes de estudar no IFMG, você estudou:")
     base_socio6 = base_socio.loc[base_socio['CODQUESTAO'] == 6]
     agrupar_questao6 = base_socio6.groupby(['RESPOSTA']).size()
     agrupar_questao6
```

```
[ ]  agrupar_questao6.plot.bar(color = 'purple')
```

```
[ ]  # Moda questão 6
     print("Antes de estudar no IFMG, você estudou:")
     moda = base_socio6.loc[base_socio['CODQUESTAO'] == 6]
     moda['RESPOSTA'].mode()[0]
```

Questão 7

```
[ ]  # Qtd de pessoas por resposta
     print("Situação do pai:")
     base_socio7 = base_socio.loc[base_socio['CODQUESTAO'] == 7]
     agrupar_questao7 = base_socio7.groupby(['RESPOSTA']).size()
     agrupar_questao7
```

```
[ ]  agrupar_questao7.plot.bar(color = 'red')
```

```
[ ]  # Moda questão 7
     print("Situação do pai:")
     moda = base_socio7.loc[base_socio['CODQUESTAO'] == 7]
     moda['RESPOSTA'].mode()[0]
```

Questão 8

```
[ ]  # Qtd de pessoas por resposta
     print("Grau de instrução do pai:")
     base_socio8 = base_socio.loc[base_socio['CODQUESTAO'] == 8]
     agrupar_questao8 = base_socio8.groupby(['RESPOSTA']).size()
     agrupar_questao8
```

```
     agrupar_questao8.plot.bar(color = 'green')
```

```
     # Moda questão 8
     print("Grau de instrução do pai:")
     moda = base_socio8.loc[base_socio['CODQUESTAO'] == 8]
     moda['RESPOSTA'].mode()[0]
```

Questão 9

```
[ ]  # Qtd de pessoas por resposta
     print("Situação da mãe:")
     base_socio9 = base_socio.loc[base_socio['CODQUESTAO'] == 9]
     agrupar_questao9 = base_socio9.groupby(['RESPOSTA']).size()
     agrupar_questao9
```

```
[ ]  agrupar_questao9.plot.bar(color = 'blue')
```

```
[ ]  # Moda questão 9
     print("Situação da mãe:")
     moda = base_socio9.loc[base_socio['CODQUESTAO'] == 9]
     moda['RESPOSTA'].mode()[0]
```

Questão 10

```
[ ]  # Qtd de pessoas por resposta
     print("Grau de instrução da mãe:")
     base_socio10 = base_socio.loc[base_socio['CODQUESTAO'] == 10]
     agrupar_questao10 = base_socio10.groupby(['RESPOSTA']).size()
     agrupar_questao10
```

```
[ ]  agrupar_questao10.plot.bar(color = 'purple')
```

```
[ ]  # Moda questão 10
     print("Grau de instrução da mãe:")
     moda = base_socio10.loc[base_socio['CODQUESTAO'] == 10]
     moda['RESPOSTA'].mode()[0]
```

Questão 11

```
[ ]  # Qtd de pessoas por resposta
     print("Você reside:")
     base_socio11 = base_socio.loc[base_socio['CODQUESTAO'] == 11]
     agrupar_questao11 = base_socio11.groupby(['RESPOSTA']).size()
     agrupar_questao11
```

```
     agrupar_questao11.plot.bar(color = 'black')
```

```
     # Moda questão 11
     print("Você reside:")
     moda = base_socio11.loc[base_socio['CODQUESTAO'] == 11]
     moda['RESPOSTA'].mode()[0]
```

Questão 12

```
[ ]  # Qtd de pessoas por resposta
     print("Residência:")
     base_socio12 = base_socio.loc[base_socio['CODQUESTAO'] == 12]
     agrupar_questao12 = base_socio12.groupby(['RESPOSTA']).size()
     agrupar_questao12
```

```
     agrupar_questao12.plot.bar(color = 'pink')
```

```
[ ]  # Moda questão 12
     print("Residência:")
     moda = base_socio12.loc[base_socio['CODQUESTAO'] == 12]
     moda['RESPOSTA'].mode()[0]
```

Questão 13

```
[ ]  # Qtd de pessoas por resposta
     print("Área de procedência:")
     base_socio13 = base_socio.loc[base_socio['CODQUESTAO'] == 13]
     agrupar_questao13 = base_socio13.groupby(['RESPOSTA']).size()
     agrupar_questao13
```

```
     agrupar_questao13.plot.bar(color = 'orange')
```

```
[ ]  # Moda questão 13
     print("Área de procedência:")
     moda = base_socio13.loc[base_socio['CODQUESTAO'] == 13]
     moda['RESPOSTA'].mode()[0]
```

Questão 14

```
[ ]  # Qtd de pessoas por resposta
     print("Renda familiar (em salários mínimos - digite somente números inteiros):")
     base_socio14 = base_socio.loc[base_socio['CODQUESTAO'] == 14]
     agrupar_questao14 = base_socio14.groupby(['RESPOSTA']).size()
     agrupar_questao14

[ ]  agrupar_questao14.plot.bar(color = 'green')

[ ]  # Média questão 14
     print("Renda familiar (em salários mínimos - digite somente números inteiros):")
     media = base_socio14.loc[base_socio['CODQUESTAO'] == 14]
     media['RESPOSTA'].mean()

[ ]  # Mínimo questão 14
     print("Renda familiar (em salários mínimos - digite somente números inteiros):")
     media = base_socio14.loc[base_socio['CODQUESTAO'] == 14]
     media['RESPOSTA'].min()

[ ]  # Máximo questão 14
     print("Renda familiar (em salários mínimos - digite somente números inteiros):")
     media = base_socio14.loc[base_socio['CODQUESTAO'] == 14]
     media['RESPOSTA'].max()
```

Questão 15

```
     # Qtd de pessoas por resposta
     print("Número de pessoas que compõem a família (inclusive você):")
     base_socio15 = base_socio.loc[base_socio['CODQUESTAO'] == 15]
     agrupar_questao15 = base_socio15.groupby(['RESPOSTA']).size()
     agrupar_questao15

     agrupar_questao15.plot.bar(color = 'red')

     # Média questão 15
     print("Número de pessoas que compõem a família (inclusive você);")
     media = base_socio15.loc[base_socio['CODQUESTAO'] == 15]
     media['RESPOSTA'].mean()

[ ]  # Máximo questão 15
     print("Número de pessoas que compõem a família (inclusive você);")
     media = base_socio15.loc[base_socio['CODQUESTAO'] == 15]
     media['RESPOSTA'].max()

[ ]  # Mínimo questão 15
     print("Número de pessoas que compõem a família (inclusive você);")
     media = base_socio15.loc[base_socio['CODQUESTAO'] == 15]
     media['RESPOSTA'].min()
```

Fonte: Elaborado pelos autores, 2023.

APÊNDICE E - Códigos utilizados para o experimento A.

Persona Geral - Indicadores de Fluxo

Desistência

Idade

```
p_idade_g = base_ifmg.loc[base_ifmg['STATUS_CURSO'].isin(['Trancado', 'Evasão', 'Desligado'])]
p_idade_g
```

```
p_idade_g['IDADE ATUAL'].describe()
```

```
p_desistencia_g = p_idade_g.loc[p_idade_g['IDADE ATUAL'] == 25]
p_desistencia_g
```

```
p_desistencia_g['COD_PESSOA'].unique()
```

Disciplinas x Pessoa

```
p_disciplina_g = p_desistencia_g.groupby(['COD_PESSOA'])['NOME_DISCIPLINA'].count()
p_disciplina_g
```

```
p_disciplina_g.describe()
```

Sexo

```
p_sexo_g = p_desistencia_g.groupby(['COD_PESSOA', 'SEXO']).count()
p_sexo_g_a = p_sexo_g.groupby(['SEXO']).size()
p_sexo_g_a
```

```
sts.mode(p_desistencia_g['SEXO'])
```

Frequência

```
p_frequencia_g = p_desistencia_g.groupby(['COD_PESSOA', 'NOME_DISCIPLINA'])['TOTAL_FALTAS'].sum()
p_frequencia_g
```

```
p_frequencia_g.describe()
```

Reprovação matérias

```
p_reprovacao_materia_g = p_desistencia_g.loc[p_desistencia_g['STATUS_DISCIPLINA'].isin(['Reprovado', 'Reprovado por frequênc
p_reprovacao_materia_g
```

```
p_reprovacao_g = p_reprovacao_materia_g.groupby(['COD_PESSOA'])['STATUS_DISCIPLINA'].count()
p_reprovacao_g
```

```
p_reprovacao_g.describe()
```

Permanência

Idade

```
[ ]  pp_idade_g = base_ifmg.loc[base_ifmg['STATUS_CURSO'].isin(['Matriculado (PD)'])]
     pp_idade_g
```

Conclusão

Idade

```
[ ]  pc_idade_g = base_ifmg.loc[base_ifmg['STATUS_CURSO'].isin(['Formado'])]
     pc_idade_g
```

Desistencia

```
[ ]  base_persona_desistencia = base_socio.loc[base_socio['CODIGO'].isin([ 47049, 52959, 56896, 161519, 165813, 190458, 78876, 92295, 11
     base_persona_desistencia
```

Permanencia

```
[ ]  base_persona_permanencia = base_socio.loc[base_socio['CODIGO'].isin([ 38903, 40148, 41501, 42290, 45860,
     base_persona_permanencia
```

Conclusão

```
[ ]  base_persona_conclusao = base_socio.loc[base_socio['CODIGO'].isin([ 40091,  40190,  45327,  45946,  5089
     base_persona_conclusao
```

Questão 6

```
[ ]  base_persona_conclusao6 = base_persona_conclusao.loc[base_persona_conclusao['CODQUESTAO'] == 6]
     agrupar_questao_conclusao6 = base_persona_conclusao6.groupby(['RESPOSTA']).size()
     agrupar_questao_conclusao6
```

```
[ ]  agrupar_questao_desistencia6.plot.bar(color = 'purple')
```

```
[ ]  moda = base_persona_conclusao.loc[base_persona_conclusao['CODQUESTAO'] == 6]
     moda['RESPOSTA'].mode()[0]
```

Questão 7

```
[ ]  base_persona_conclusao7 = base_persona_conclusao.loc[base_persona_conclusao['CODQUESTAO'] == 7]
     agrupar_questao_conclusao7 = base_persona_conclusao7.groupby(['RESPOSTA']).size()
     agrupar_questao_conclusao7
```

```
[ ]  agrupar_questao_conclusao7.plot.bar(color = 'red')
```

```
[ ]  moda = base_persona_conclusao.loc[base_persona_conclusao['CODQUESTAO'] == 7]
     moda['RESPOSTA'].mode()[0]
```

Questão 8

```
[ ]  base_persona_conclusao8 = base_persona_conclusao.loc[base_persona_conclusao['CODQUESTAO'] == 8]
     agrupar_questao_conclusao8 = base_persona_conclusao8.groupby(['RESPOSTA']).size()
     agrupar_questao_conclusao8
```

```
[ ]  agrupar_questao_conclusao8.plot.bar(color = 'green')
```

```
[ ]  moda = base_persona_conclusao.loc[base_persona_conclusao['CODQUESTAO'] == 8]
     moda['RESPOSTA'].mode()[0]
```

Questão 9

```
[ ]  base_persona_conclusao9 = base_persona_conclusao.loc[base_persona_conclusao['CODQUESTAO'] == 9]
     agrupar_questao_conclusao9 = base_persona_conclusao9.groupby(['RESPOSTA']).size()
     agrupar_questao_conclusao9
```

```
[ ]  agrupar_questao_conclusao9.plot.bar(color = 'blue')
```

```
[ ]  moda = base_persona_conclusao.loc[base_persona_conclusao['CODQUESTAO'] == 9]
     moda['RESPOSTA'].mode()[0]
```

Questão 10

```
[ ] base_persona_conclusao10 = base_persona_conclusao.loc[base_persona_conclusao['CODQUESTAO'] == 10]
    agrupar_questao_conclusao10 = base_persona_conclusao10.groupby(['RESPOSTA']).size()
    agrupar_questao_conclusao10
```

```
[ ] agrupar_questao_conclusao10.plot.bar(color = 'purple')
```

```
[ ] moda = base_persona_conclusao.loc[base_persona_conclusao['CODQUESTAO'] == 10]
    moda['RESPOSTA'].mode()[0]
```

Questão 11

```
[ ] base_persona_conclusao11 = base_persona_conclusao.loc[base_persona_conclusao['CODQUESTAO'] == 11]
    agrupar_questao_conclusao11 = base_persona_conclusao11.groupby(['RESPOSTA']).size()
    agrupar_questao_conclusao11
```

```
[ ] agrupar_questao_conclusao11.plot.bar(color = 'black')
```

```
[ ] moda = base_persona_conclusao.loc[base_persona_conclusao['CODQUESTAO'] == 11]
    moda['RESPOSTA'].mode()[0]
```

Questão 12

```
[ ] base_persona_conclusao12 = base_persona_conclusao.loc[base_persona_conclusao['CODQUESTAO'] == 12]
    agrupar_questao_conclusao12 = base_persona_conclusao12.groupby(['RESPOSTA']).size()
    agrupar_questao_conclusao12
```

```
[ ] agrupar_questao_conclusao12.plot.bar(color = 'pink')
```

```
[ ] moda = base_persona_conclusao.loc[base_persona_conclusao['CODQUESTAO'] == 12]
    moda['RESPOSTA'].mode()[0]
```

Questão 13

```
[ ] base_persona_conclusao13 = base_persona_conclusao.loc[base_persona_conclusao['CODQUESTAO'] == 13]
    agrupar_questao_conclusao13 = base_persona_conclusao13.groupby(['RESPOSTA']).size()
    agrupar_questao_conclusao13
```

```
[ ] agrupar_questao_conclusao13.plot.bar(color = 'orange')
```

```
[ ] moda = base_persona_conclusao.loc[base_persona_conclusao['CODQUESTAO'] == 13]
    moda['RESPOSTA'].mode()[0]
```

Dados Numéricos

```
q_permanencia = base_persona_conclusao.groupby(['CODIGO', 'QUESTAO', 'RESPOSTA']).count()
agrupar_questao_conclusao = q_permanencia.groupby(['QUESTAO','RESPOSTA']).size()
agrupar_questao_conclusao
```

Questão 14

```
base_persona_conclusao14 = base_persona_conclusao.loc[base_persona_conclusao['CODQUESTAO'] == 14]
agrupar_questao_conclusao14 = base_persona_conclusao14.groupby(['RESPOSTA']).size()
agrupar_questao_conclusao14
```

```
agrupar_questao_conclusao14.plot.bar(color = 'green')
```

```
media = base_persona_conclusao.loc[base_persona_conclusao['CODQUESTAO'] == 14]
media['RESPOSTA'].mean()
```

```
media = base_persona_conclusao.loc[base_persona_conclusao['CODQUESTAO'] == 14]
media['RESPOSTA'].min()
```

```
media = base_persona_conclusao.loc[base_persona_conclusao['CODQUESTAO'] == 14]
media['RESPOSTA'].max()
```

Questão 15

```
base_persona_conclusao15 = base_persona_conclusao.loc[base_persona_conclusao['CODQUESTAO'] == 15]
agrupar_questao_conclusao15 = base_persona_conclusao15.groupby(['RESPOSTA']).size()
agrupar_questao_conclusao15
```

```
agrupar_questao_conclusao15.plot.bar(color = 'red')
```

```
media = base_persona_conclusao.loc[base_persona_conclusao['CODQUESTAO'] == 15]
media['RESPOSTA'].mean()
```

```
media = base_persona_conclusao.loc[base_persona_conclusao['CODQUESTAO'] == 15]
media['RESPOSTA'].max()
```

```
media = base_persona_conclusao.loc[base_persona_conclusao['CODQUESTAO'] == 15]
media['RESPOSTA'].min()
```

Fonte: Elaborado pelos autores, 2023.

APÊNDICE F - Códigos utilizados para o experimento B.

Desistência

```
[ ]  base_desistencia = base_ifmg[base_ifmg['STATUS_CURSO'].isin(['Trancado', 'Evasão', 'Desligado'])]
```

```
[ ]  ano = base_desistencia.groupby(['COD_PESSOA', 'ANO']).count()
     agrupar_ano = ano.groupby(['ANO']).size()
     agrupar_ano
```

```
[ ]  agrupar_ano.plot.bar(color = 'blue')
```

```
[ ]
     ano = base_desistencia.groupby(['COD_PESSOA', 'CODPERLET']).count()
     agrupar_perlet = ano.groupby(['CODPERLET']).size()
     agrupar_perlet
```

```
[ ]  agrupar_perlet.plot.bar(color = 'blue')
```

Antes da pandemia

```
[ ]  d_ap = base_desistencia[base_desistencia['ANO'] <= 2019]
```

Persona - Antes da Pandemia(Evasão)

Idade

```
[ ]  d_ap['IDADE ATUAL'].describe()
```

```
[ ]  d_desistencia_a_p = d_ap.loc[d_ap['IDADE ATUAL'] == 26]
     d_desistencia_a_p
```

```
[ ]  d_desistencia_a_p['COD_PESSOA'].unique()
```

Disciplina x Pessoa

```
[ ]  d_disciplina_ap = d_desistencia_a_p.groupby(['COD_PESSOA'])['NOME_DISCIPLINA'].count()
     d_disciplina_ap
```

```
     d_disciplina_ap.describe()
```

Sexo

```
[ ]  d_sexo_ap = d_desistencia_a_p.groupby(['COD_PESSOA', 'SEXO']).count()
     d_sexo_ap = d_sexo_ap.groupby(['SEXO']).size()
     d_sexo_ap
```

```
[ ]  sts.mode(d_desistencia_a_p['SEXO'])
```

Frequência

```
[ ]  d_frequencia_a_p = d_desistencia_a_p.groupby(['COD_PESSOA', 'NOME_DISCIPLINA'])['TOTAL_FALTAS'].sum()
     d_frequencia_a_p
```

```
[ ]  d_frequencia_a_p.describe()
```

Reprovação matéria

```
[ ]  d_reprovacao_materia_a_p = d_desistencia_a_p.loc[d_desistencia_a_p['STATUS_DISCIPLINA'].isin(['Reprovado', 'Reprovado por f
     d_reprovacao_materia_a_p
```

```
[ ]  d_reprovacao_a_p = d_reprovacao_materia_a_p.groupby(['COD_PESSOA'])['STATUS_DISCIPLINA'].count()
     d_reprovacao_a_p
```

```
[ ]  d_reprovacao_a_p.describe()
```

Durante a pandemia

```
[ ]  base_desistencia_aux = base_desistencia[base_desistencia['ANO'] >= 2020]
     base_desistencia_d_p = base_desistencia_aux[base_desistencia_aux['ANO'] <= 2021]
```

Pos pandemia

```
[ ]  pp = base_desistencia[base_desistencia['ANO'] == 2022]
```

Conclusão(formados)

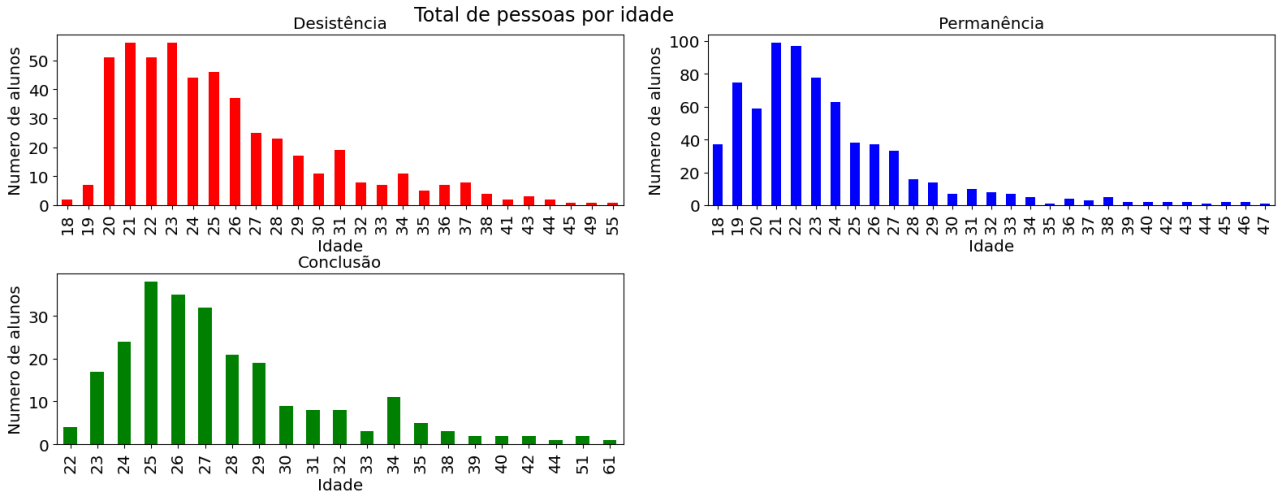
```
[ ]  base_formados = base_ifmg[base_ifmg['STATUS_CURSO'] == 'Formado']

[ ]  ano_c = base_formados.groupby(['COD_PESSOA', 'ANO']).count()
     agrupar_ano_c = ano.groupby(['ANO']).size()
     agrupar_ano_c

     agrupar_ano_c.plot.bar(color = 'blue')

     ano_c = base_formados.groupby(['COD_PESSOA', 'CODPERLET']).count()
     agrupar_perlet_c = ano_c.groupby(['CODPERLET']).size()
     agrupar_perlet_c

[ ]  agrupar_perlet.plot.bar(color = 'blue')
```

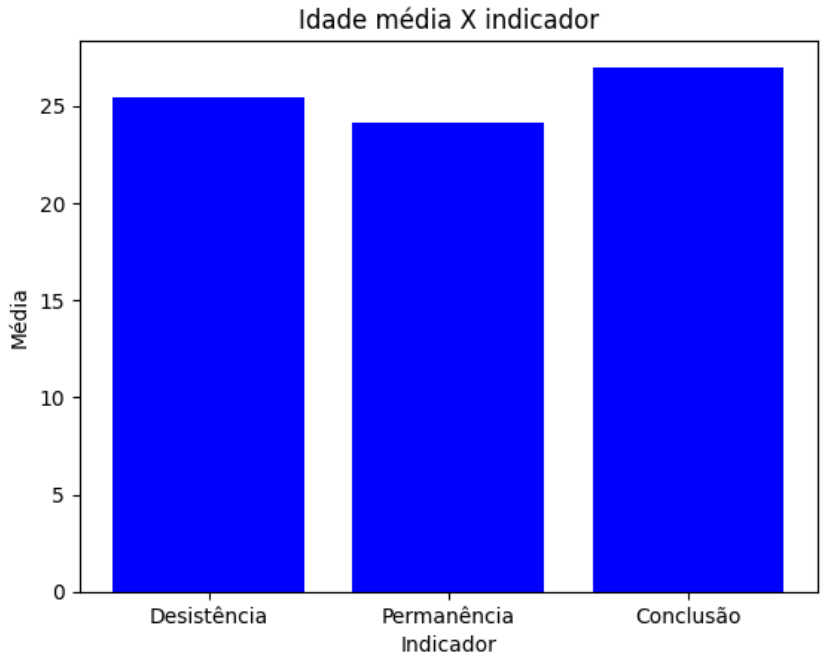
Permanência (Matriculados)

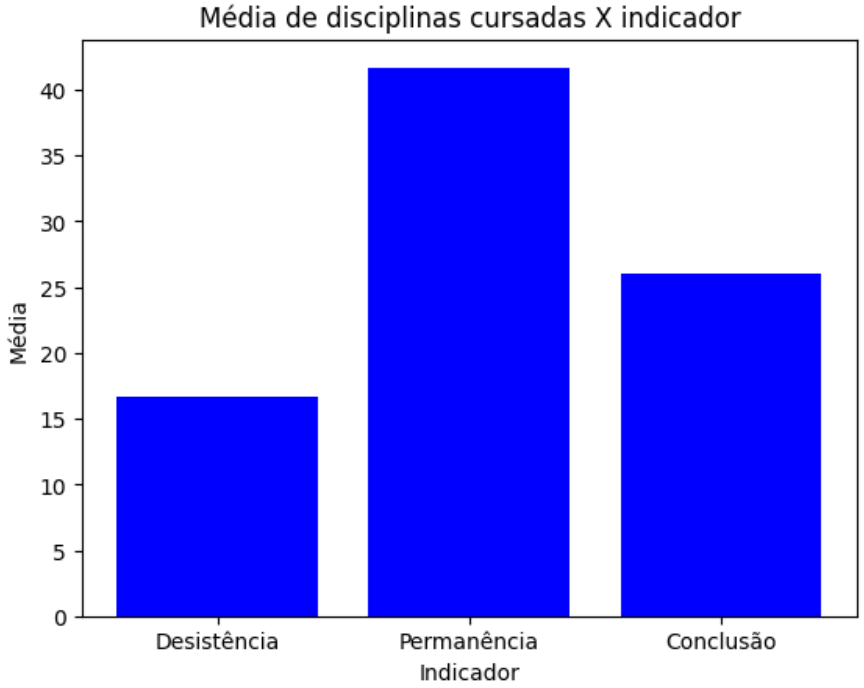
```
[ ]  base_matriculados = base_ifmg[base_ifmg['STATUS_CURSO'] == 'Matriculado (PD)']

     ano_m = base_matriculados.groupby(['COD_PESSOA', 'ANO']).count()
     agrupar_ano_m = ano_m.groupby(['ANO']).size()
     agrupar_ano_m

     agrupar_ano_m.plot.bar(color = 'blue')

[ ]  ##agrupar por perlet
     ano_ma = base_matriculados.groupby(['COD_PESSOA', 'CODPERLET']).count()
     agrupar_perlet_m = ano_ma.groupby(['CODPERLET']).size()
     agrupar_perlet_m

     agrupar_perlet_m.plot.bar(color = 'blue')
```

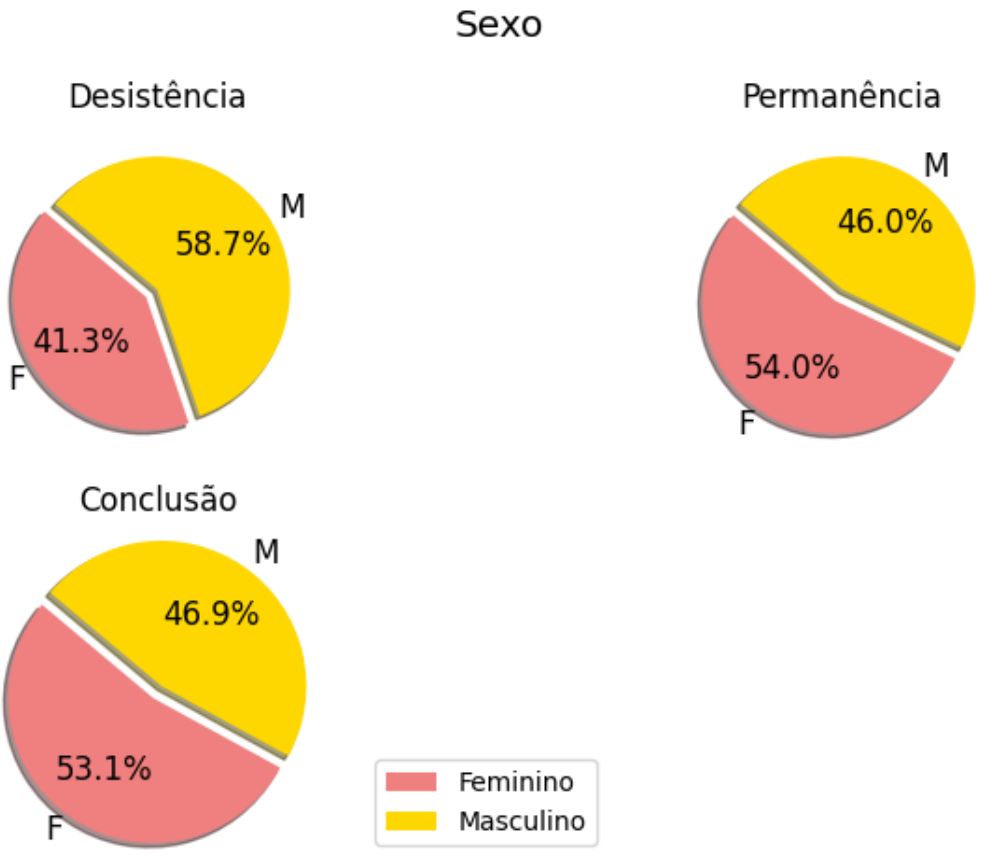
Fonte: Elaborado pelos autores, 2023.

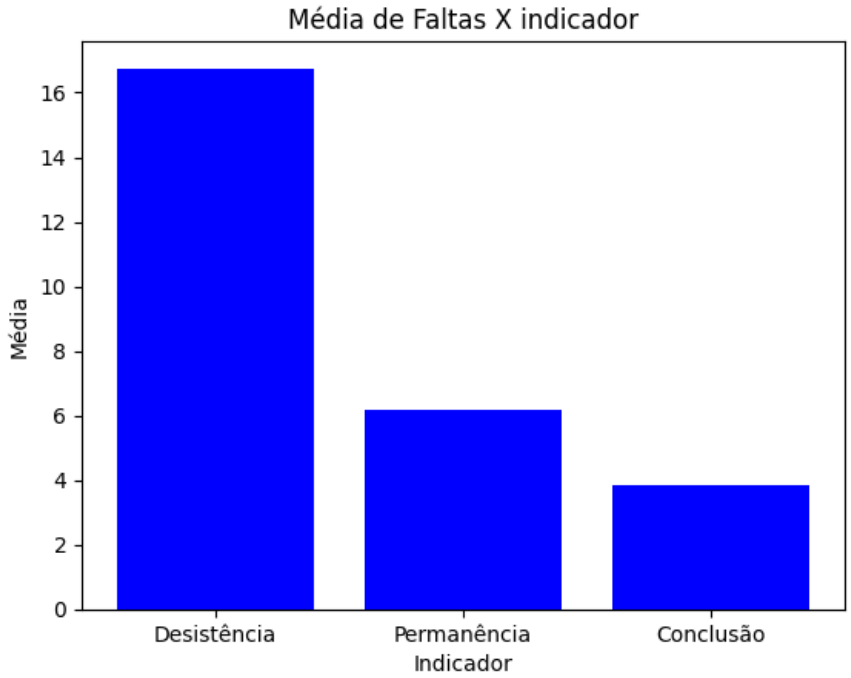
APÊNDICE G - Gráficos do Experimento A.

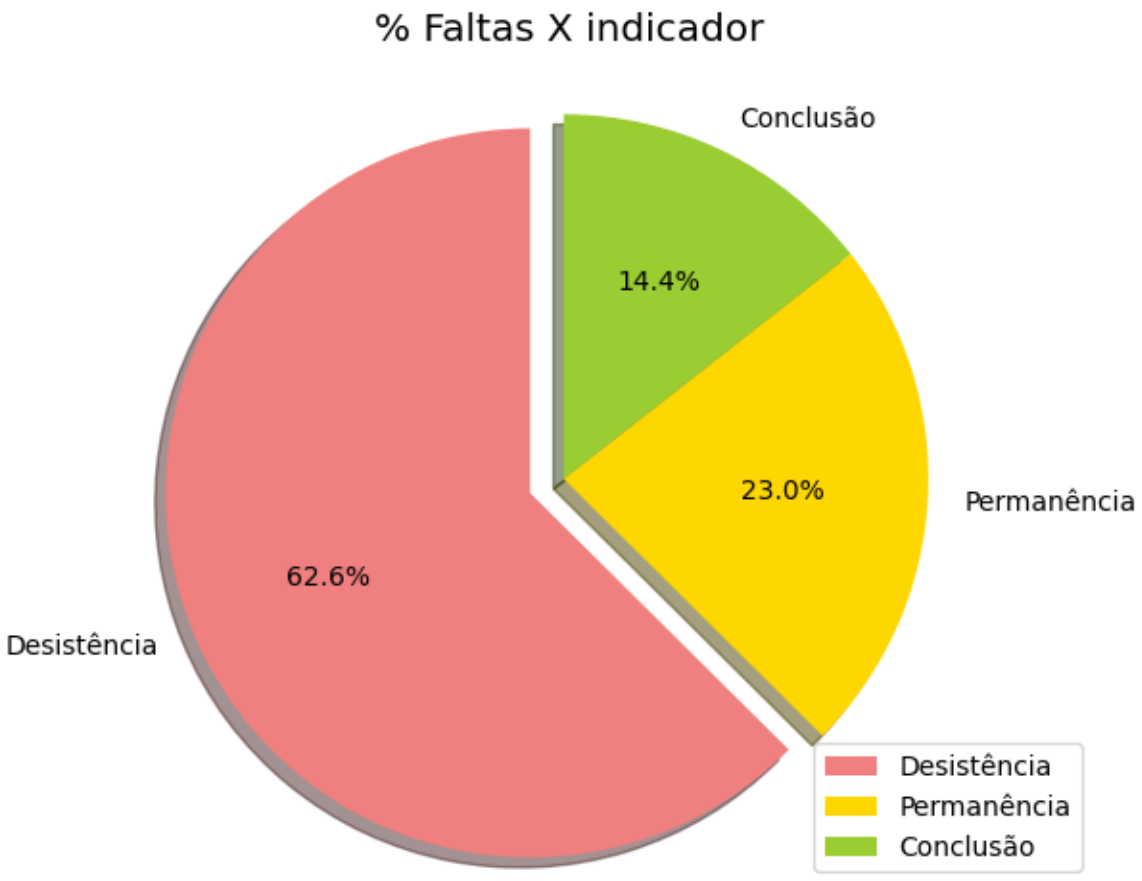


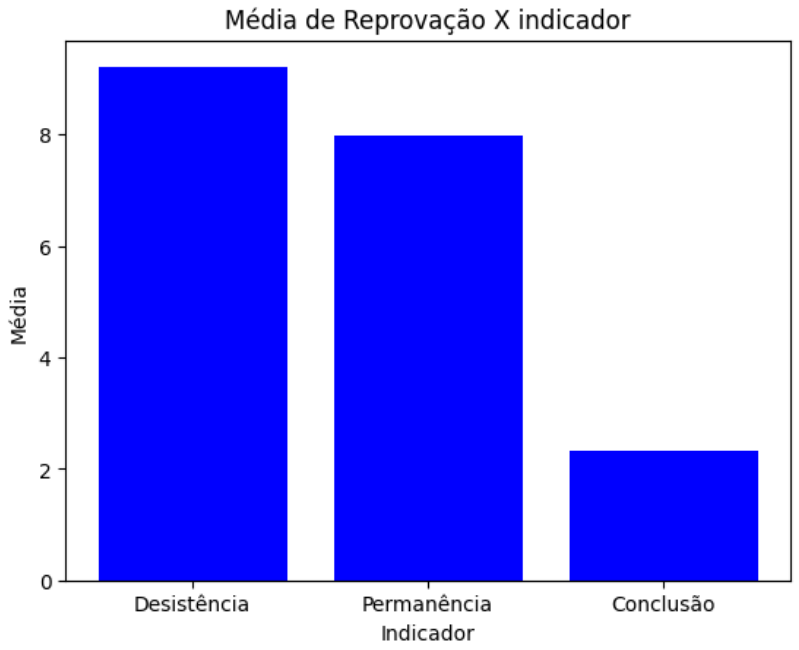


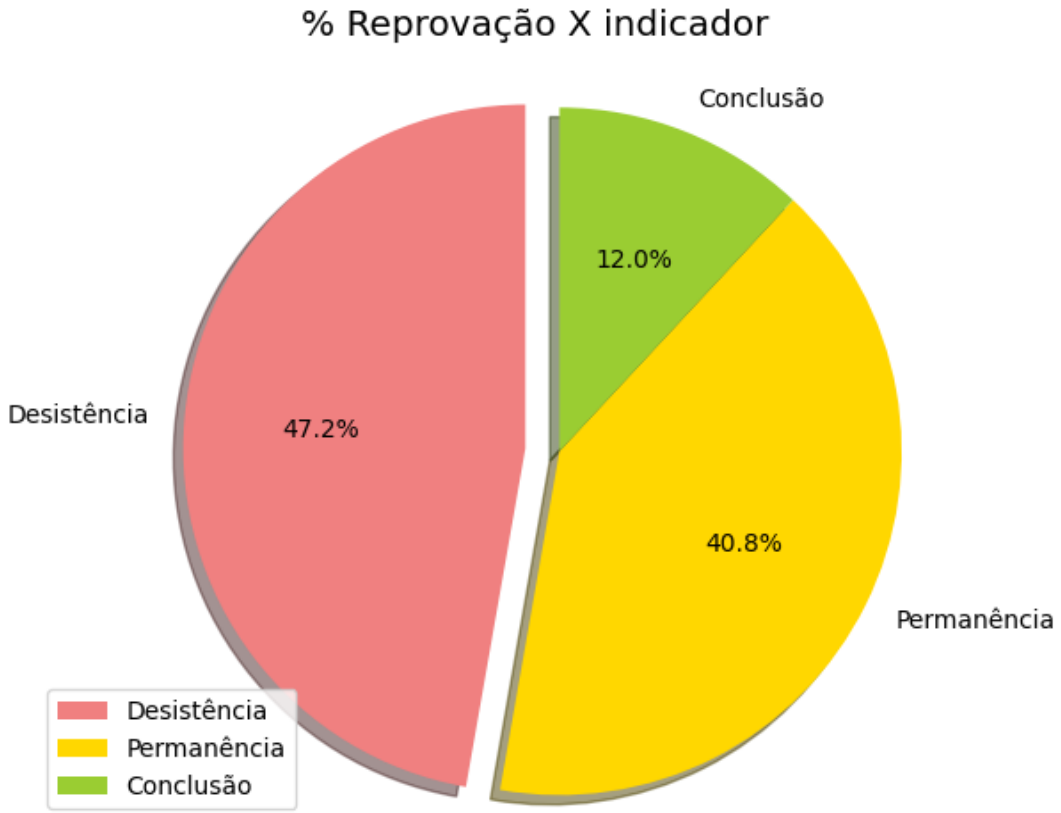

Média de disciplinas cursadas X indicador
Média
0
5
10
15
20
25
30
35
40
Desistência
Permanência
Conclusão
Indicador

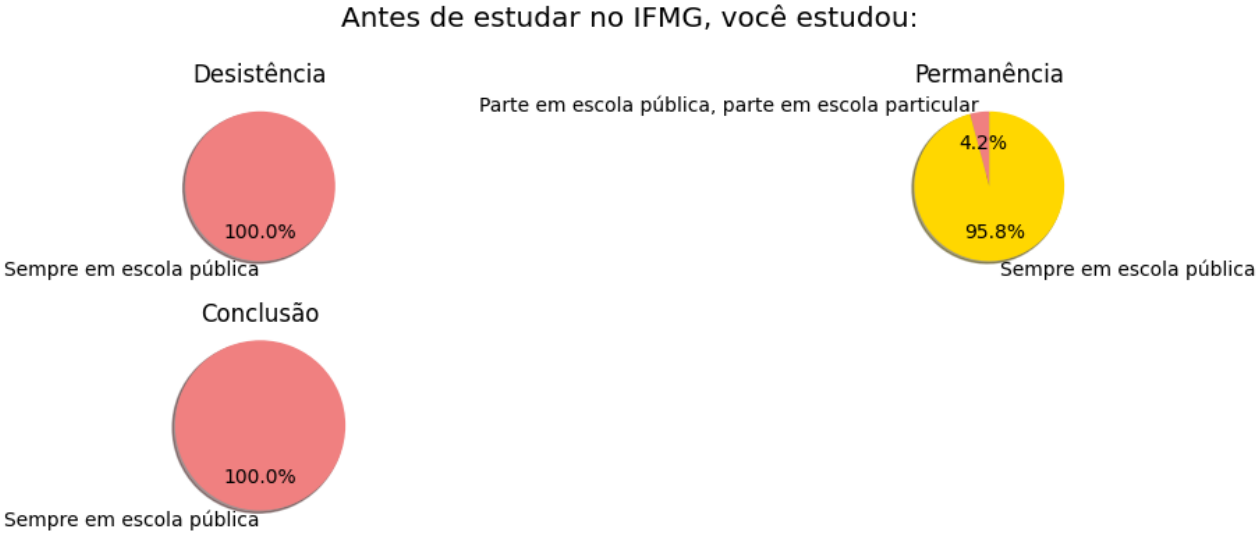

Sexo
Desistência
M
58.7%
41.3%
F
Permanência
M
46.0%
54.0%
F
Conclusão
M
46.9%
53.1%
F
Feminino
Masculino


Média de Faltas X indicador
Média
16
14
12
10
8
6
4
2
0
Desistência
Permanência
Conclusão
Indicador

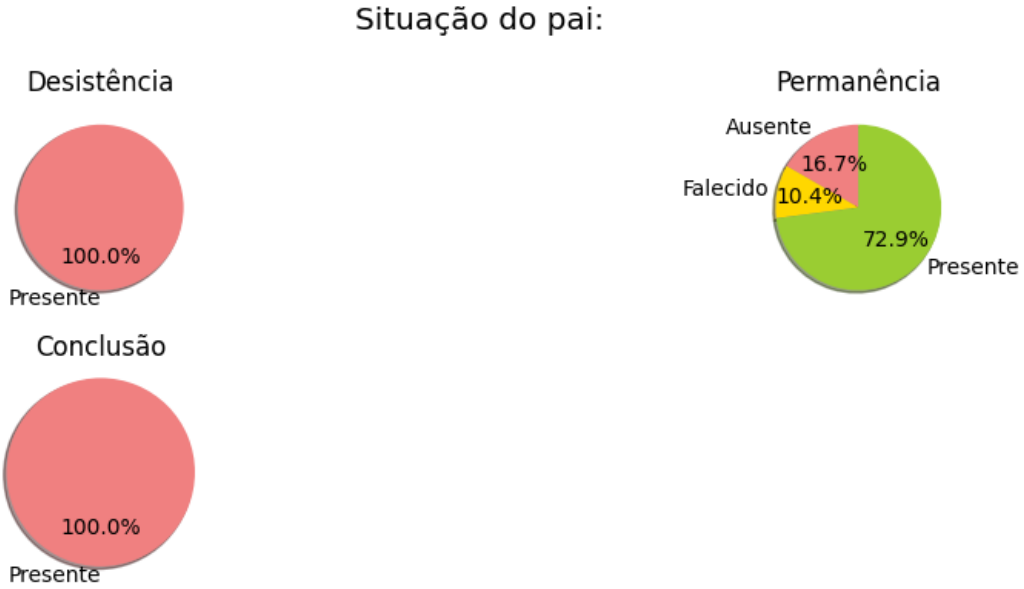

% Faltas X indicador
Conclusão
14.4%
23.0%
Permanência
62.6%
Desistência
Desistência
Permanência
Conclusão


Média de Reprovação X indicador
Média
0
2
4
6
8
Desistência
Permanência
Conclusão
Indicador

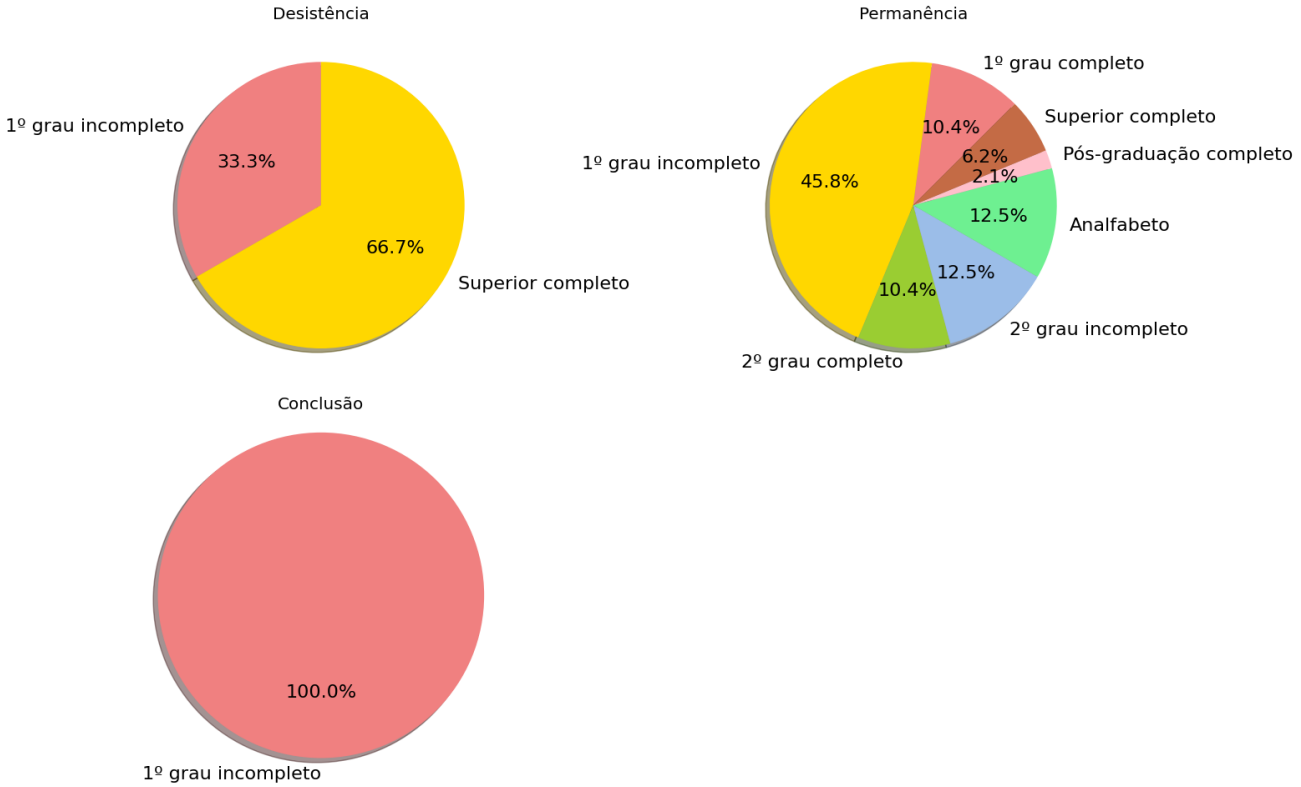

% Reprovação X indicador
Conclusão
12.0%
Desistência
47.2%
40.8%
Permanência
Desistência
Permanência
Conclusão


Antes de estudar no IFMG, você estudou:
Desistência
100.0%
Sempre em escola pública
Permanência
Parte em escola pública, parte em escola particular
4.2%
95.8%
Sempre em escola pública
Conclusão
100.0%
Sempre em escola pública

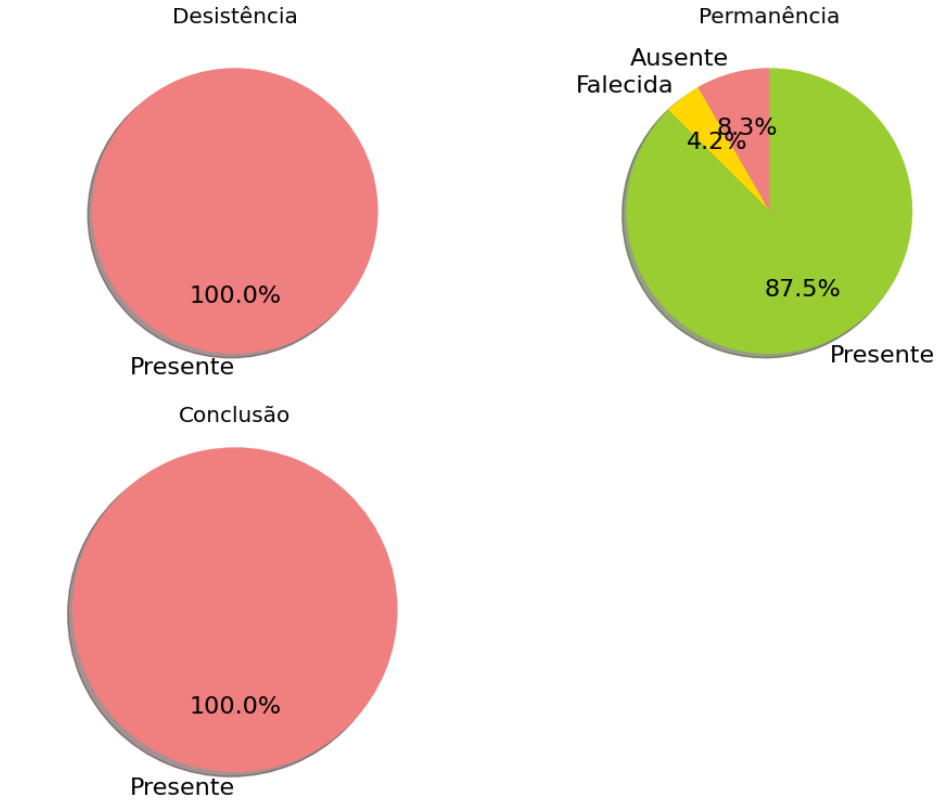
.


Situação do pai:
Desistência
100.0%
Presente
Permanência
Ausente
16.7%
Falecido
10.4%
72.9%
Presente
Conclusão
100.0%
Presente

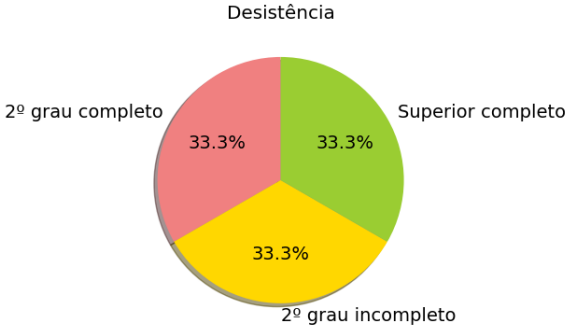
Grau de instrução do pai:

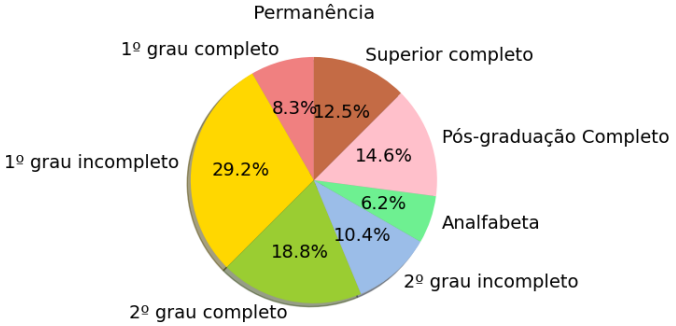


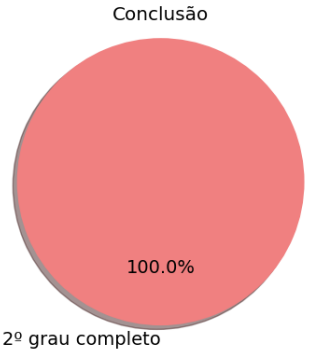
Situação da mãe:



Grau de instrução da mãe:





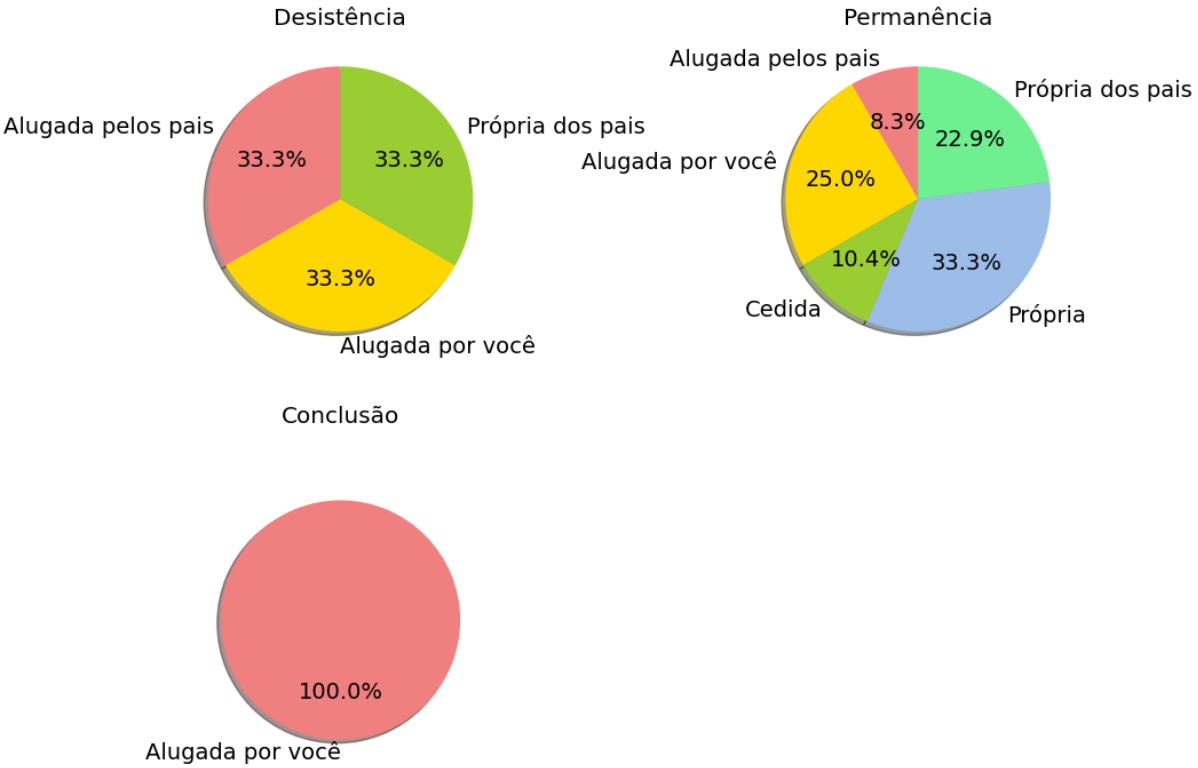


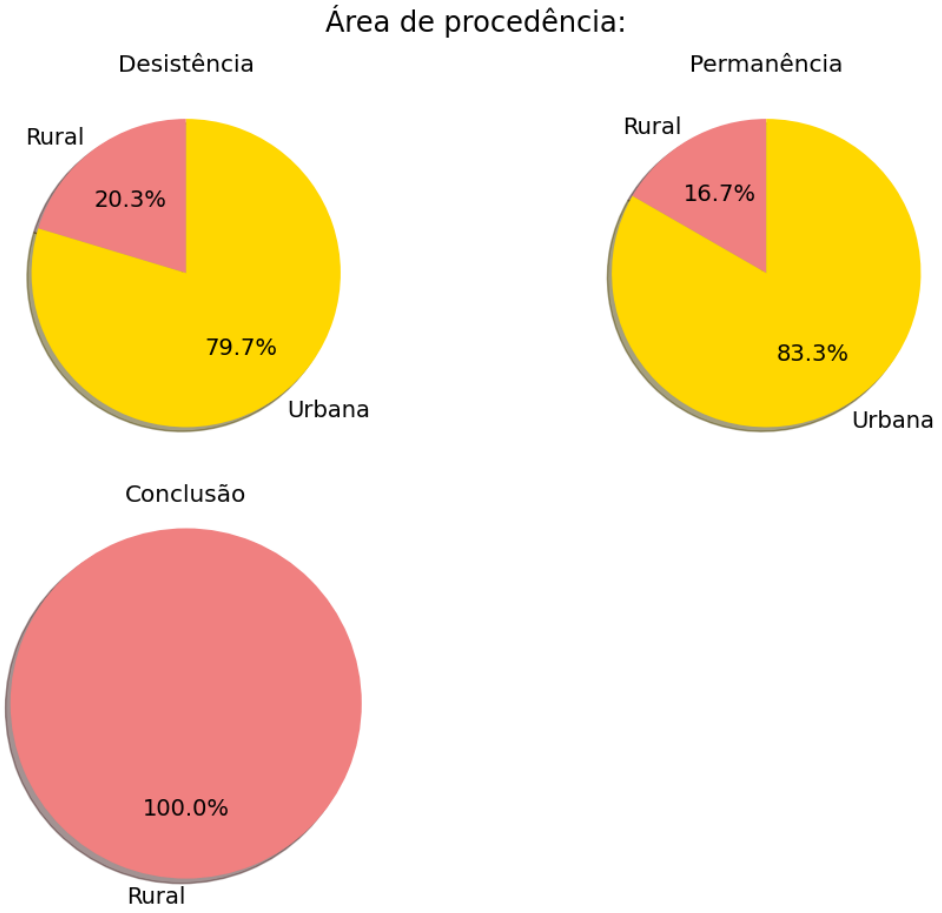
Desistência
Com os pais
33.3%
33.3%
33.3%
Outro (Clique no botão "Inserir comentário" à direita para especificar)
Em casa de estudantes

Você reside:
Conclusão
100.0%
Outro (Clique no botão "Inserir comentário" à direita para especificar)

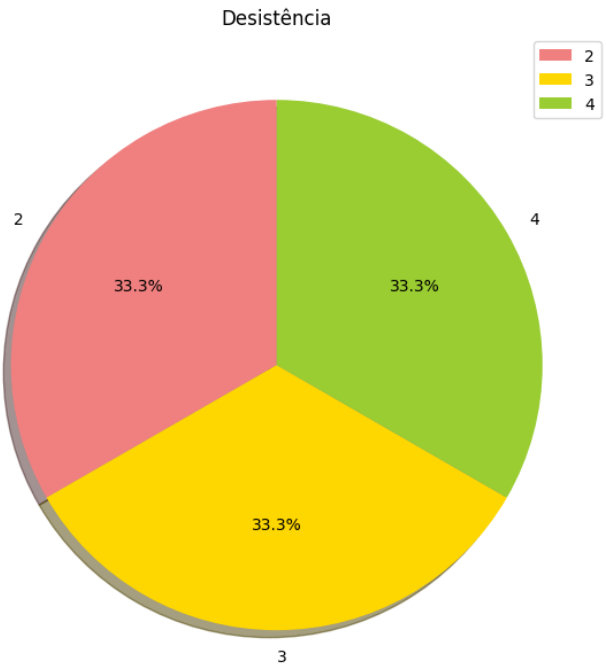
Permanência
Com a mãe
Com amigos
Com o cônjuge
Com o pai
Com os pais
8.3%
4.2%
4.2%
2.1%
20.8%
2.1%
29.2%
29.2%
Outro (Clique no botão "Inserir comentário" à direita para especificar)
Em pensionato
Em casa de estudantes

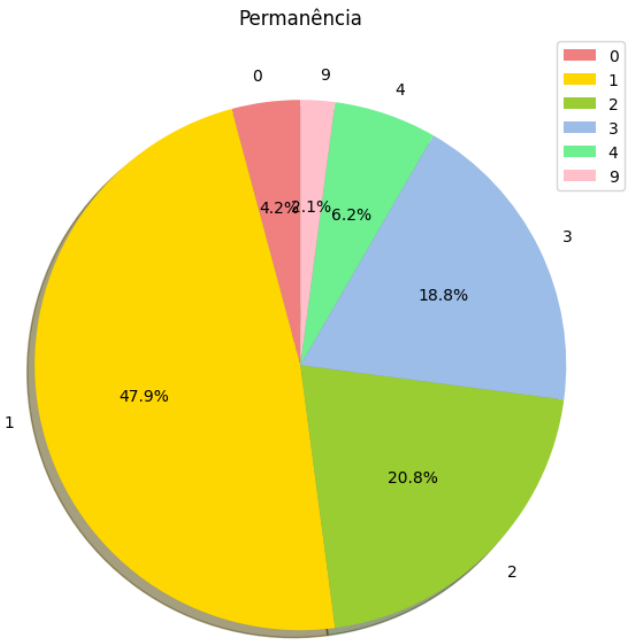
Residência:

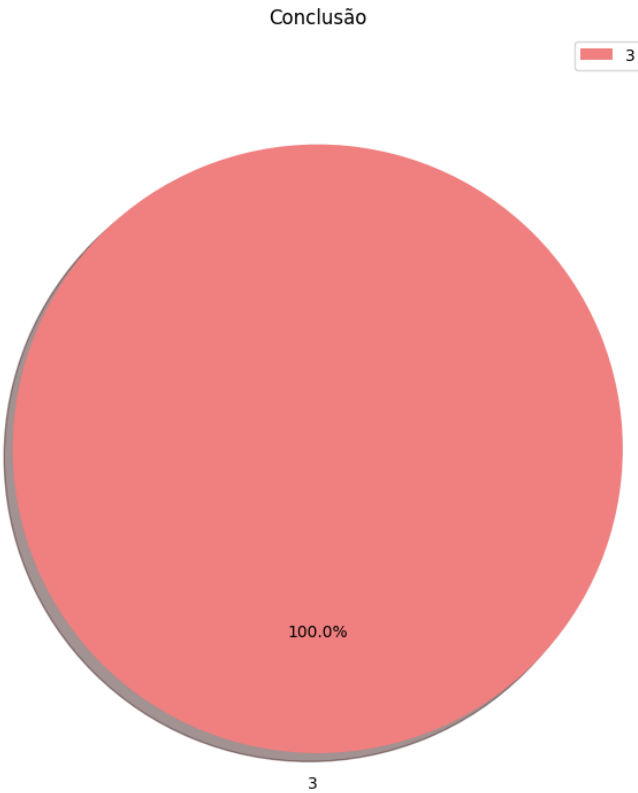

Desistência
Alugada pelos pais
Própria dos pais
33.3%
33.3%
33.3%
Alugada por você
Permanência
Alugada pelos pais
Própria dos pais
Alugada por você
8.3%
22.9%
25.0%
10.4%
33.3%
Cedida
Própria
Conclusão
100.0%
Alugada por você


Área de procedência:
Desistência
Rural
20.3%
79.7%
Urbana
Permanência
Rural
16.7%
83.3%
Urbana
Conclusão
100.0%
Rural

Renda familiar (em salários mínimos - digite somente números inteiros):

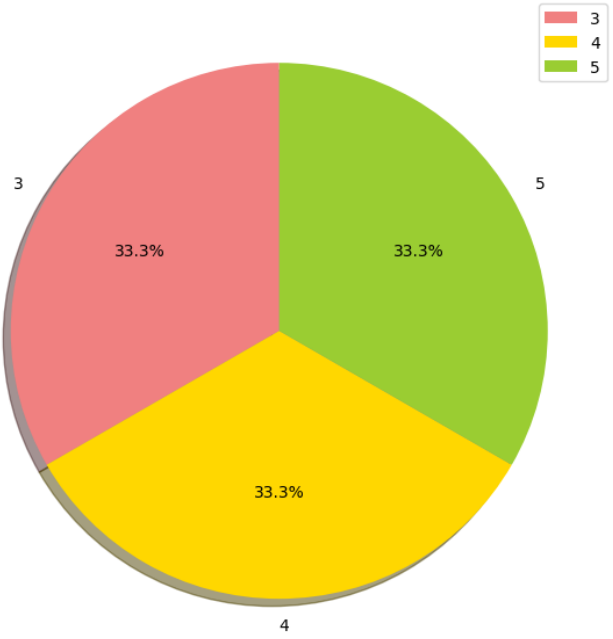




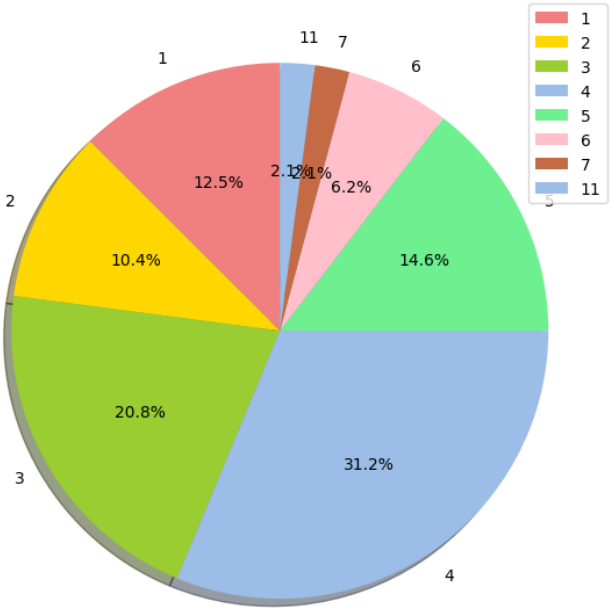


Número de pessoas que compõem a família (inclusive você):

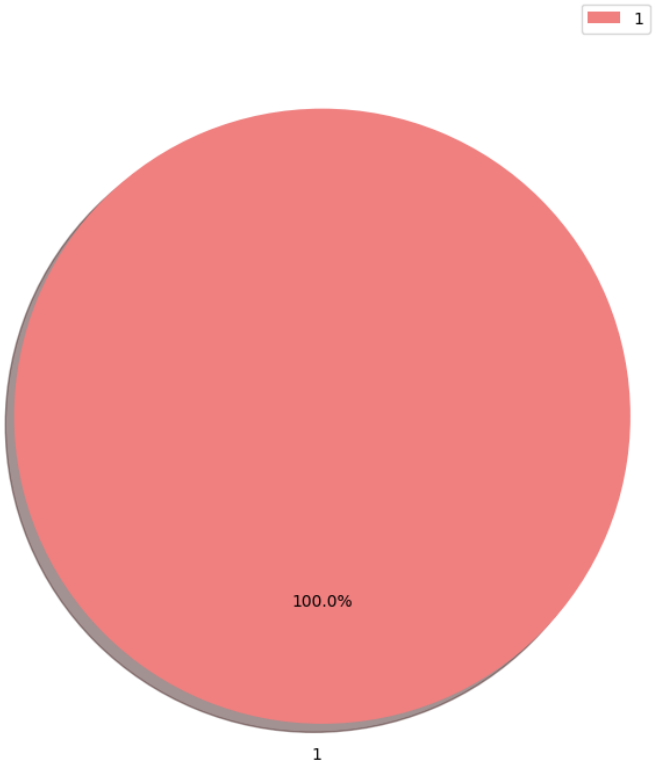
Desistência



Permanência

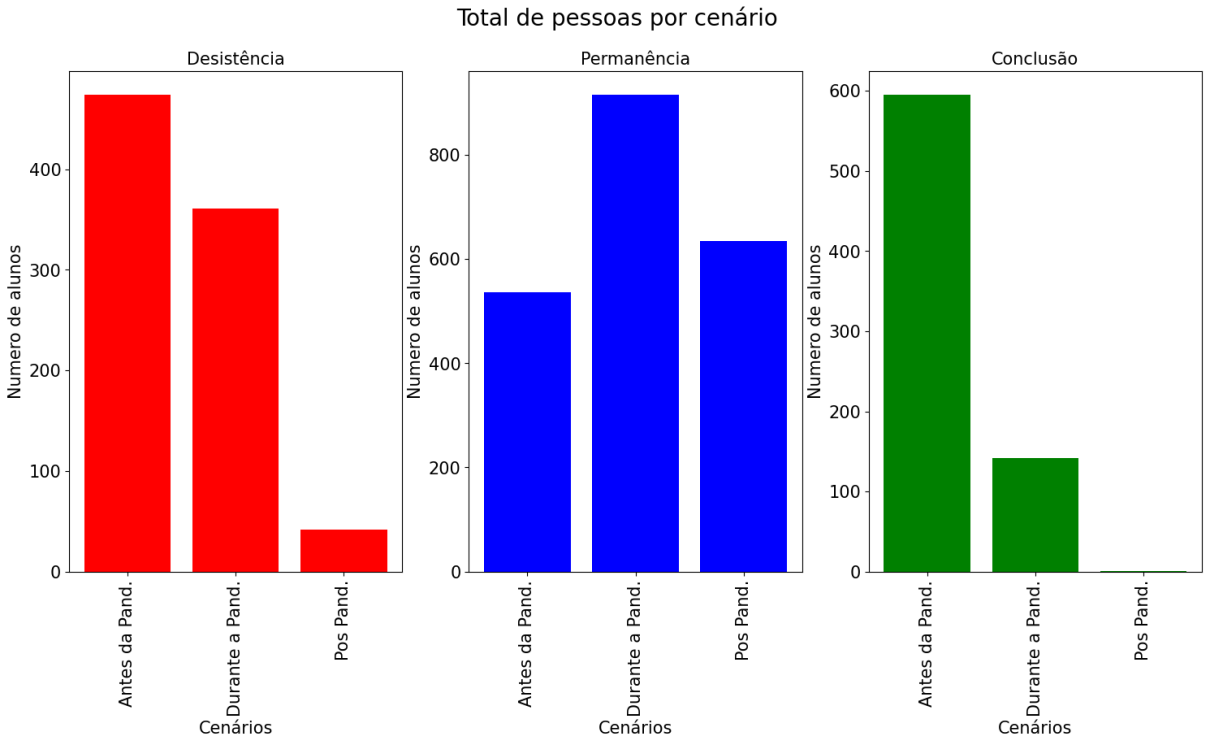


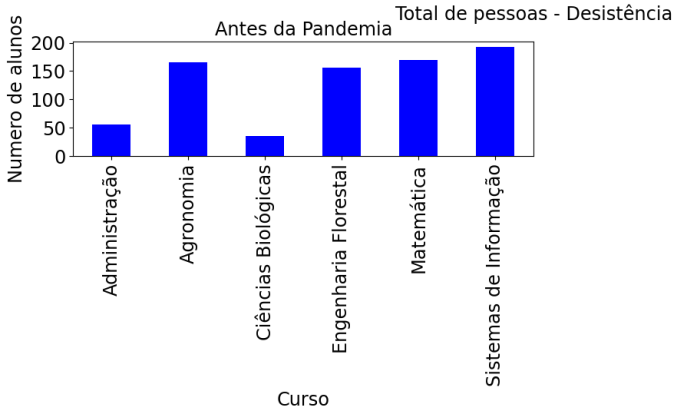
Conclusão

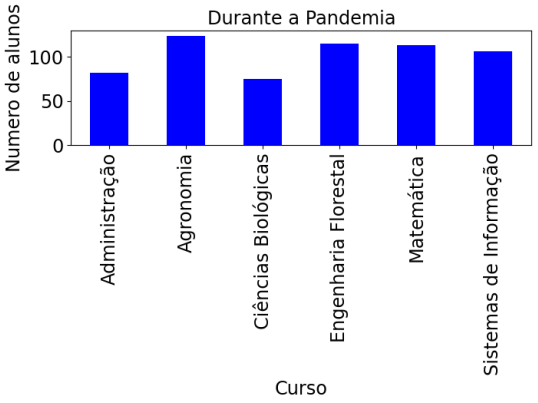


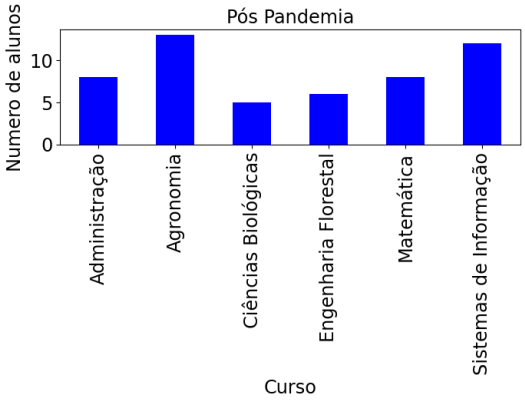
Fonte: Elaborado pelos autores, 2023.

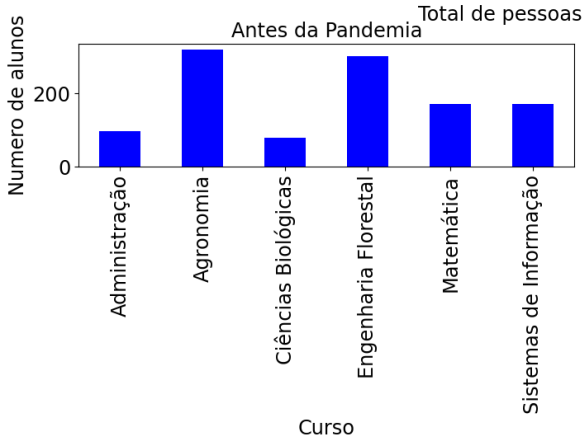
APÊNDICE H - Gráficos do Experimento B.

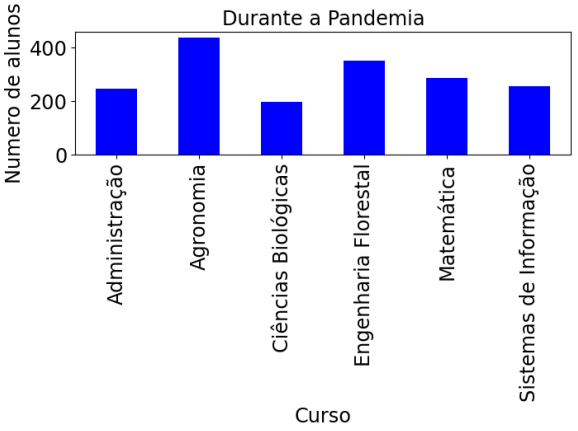


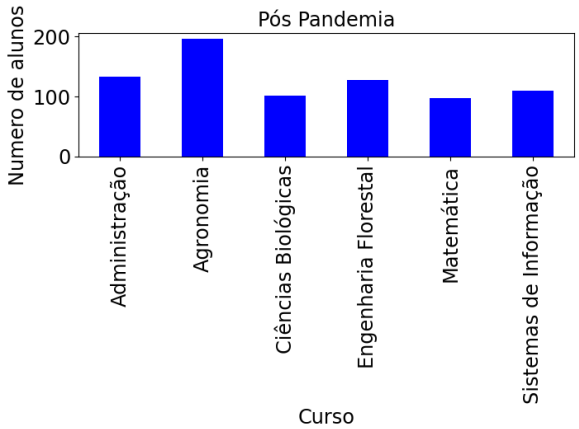

Total de pessoas - Desistência
Antes da Pandemia
Numero de alunos
200
150
100
50
0
Administração
Agronomia
Ciências Biológicas
Engenharia Florestal
Matemática
Sistemas de Informação
Curso

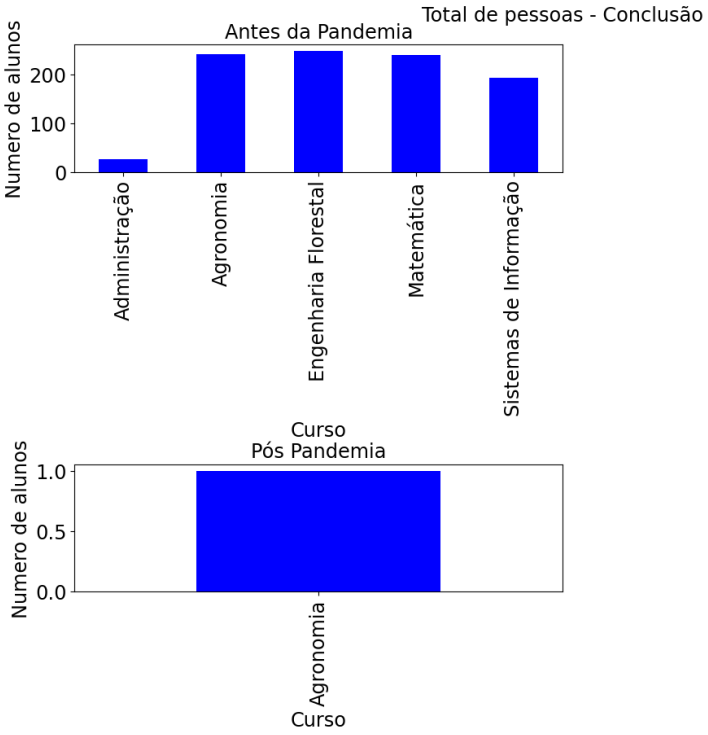

Durante a Pandemia
Numero de alunos
100
50
0
Administração
Agronomia
Ciências Biológicas
Engenharia Florestal
Matemática
Sistemas de Informação
Curso

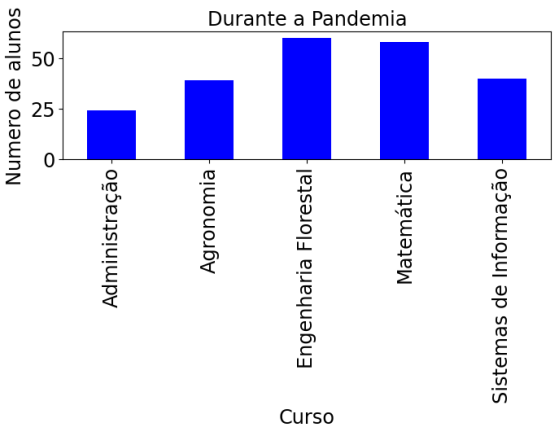

Pós Pandemia
Numero de alunos
10
5
0
Administração
Agronomia
Ciências Biológicas
Engenharia Florestal
Matemática
Sistemas de Informação
Curso


Total de pessoas - Permanência
Antes da Pandemia
Numero de alunos
200
0
Administração
Agronomia
Ciências Biológicas
Engenharia Florestal
Matemática
Sistemas de Informação
Curso

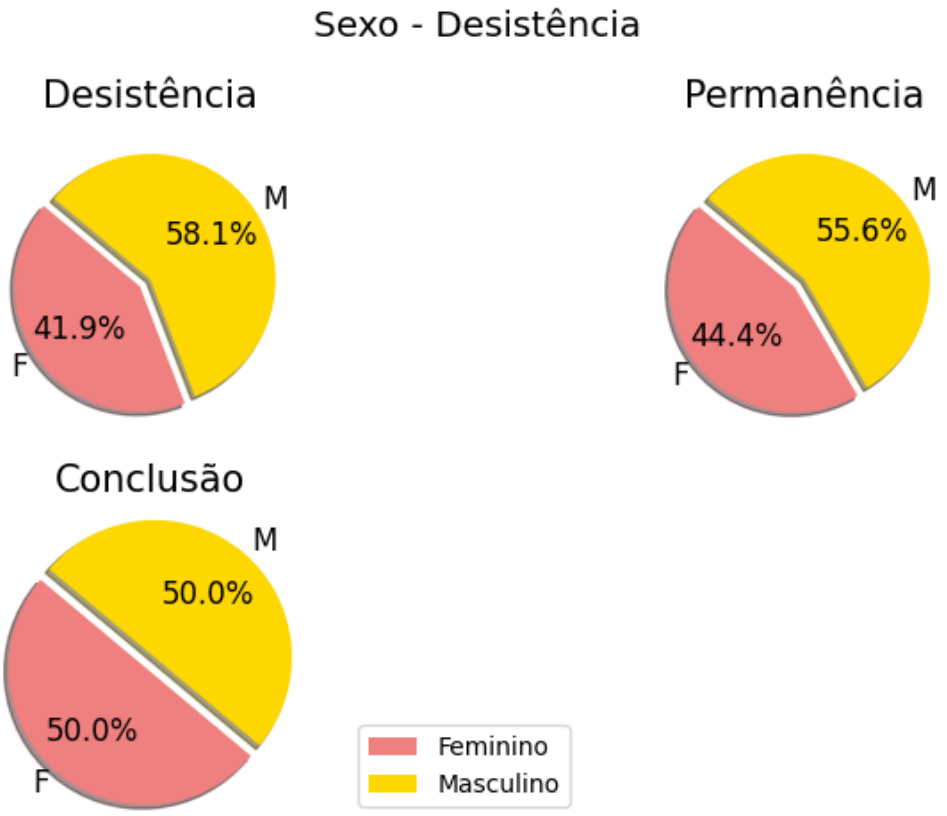

Durante a Pandemia
Numero de alunos
400
200
0
Administração
Agronomia
Ciências Biológicas
Engenharia Florestal
Matemática
Sistemas de Informação
Curso

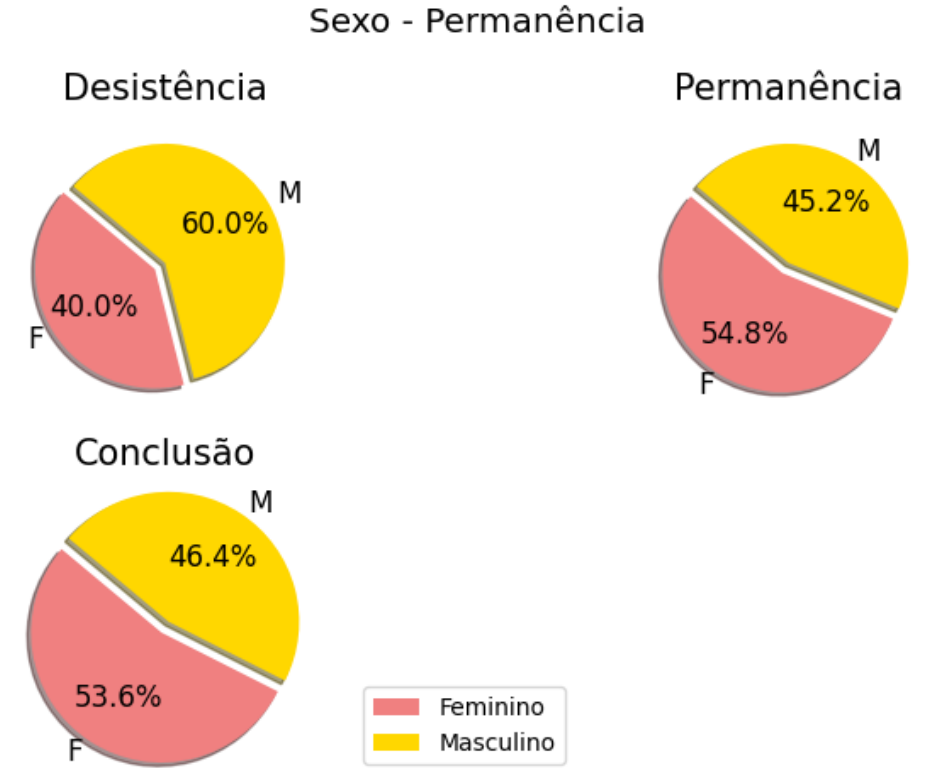

Pós Pandemia
Numero de alunos
200
100
0
Administração
Agronomia
Ciências Biológicas
Engenharia Florestal
Matemática
Sistemas de Informação
Curso

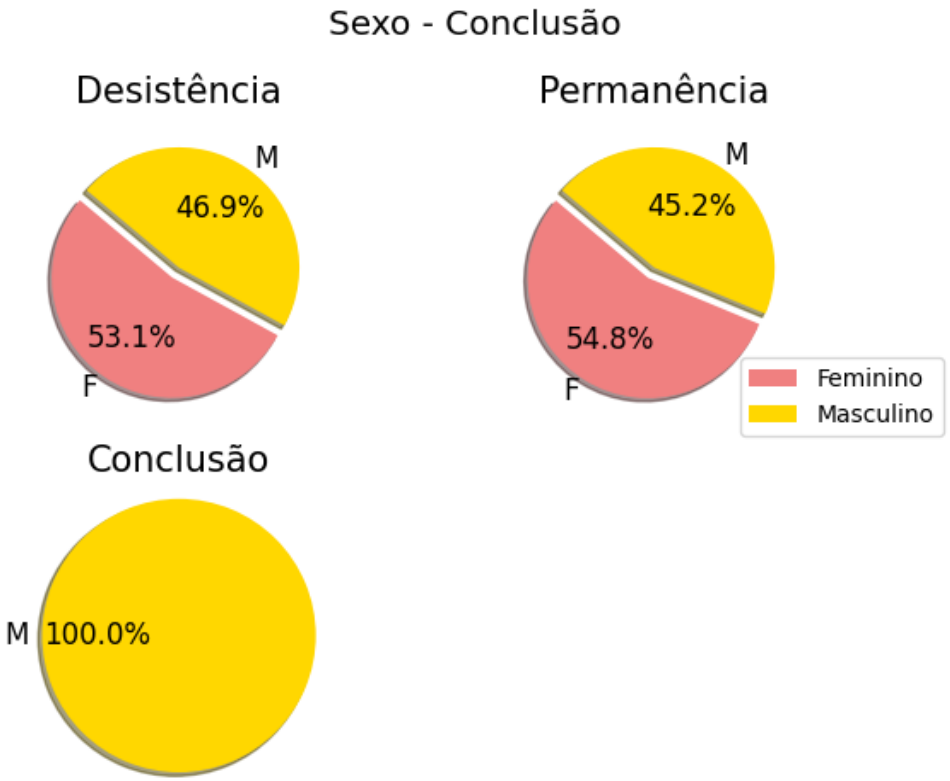

Total de pessoas - Conclusão
Antes da Pandemia
Numero de alunos
0
100
200
Administração
Agronomia
Engenharia Florestal
Matemática
Sistemas de Informação
Curso


Durante a Pandemia
Numero de alunos
0
25
50
Administração
Agronomia
Engenharia Florestal
Matemática
Sistemas de Informação
Curso

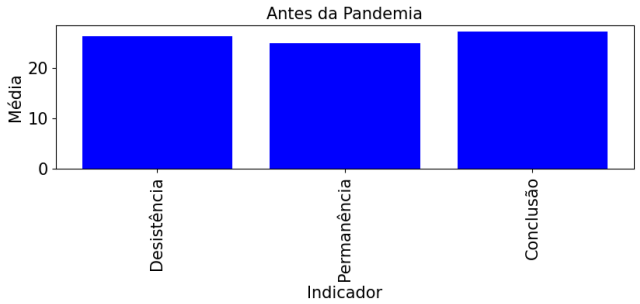
Pós Pandemia
Numero de alunos
0.0
0.5
1.0
Agronomia
Curso

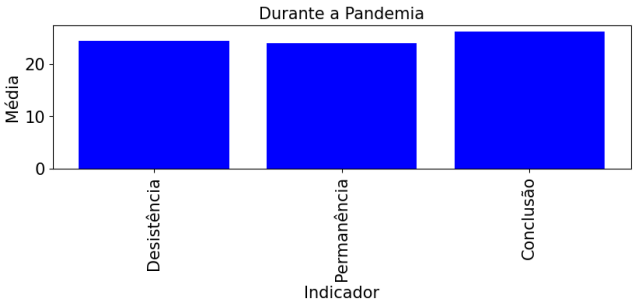

Sexo - Desistência
Desistência
M
58.1%
41.9%
F
Permanência
M
55.6%
44.4%
F
Conclusão
M
50.0%
50.0%
F
Feminino
Masculino

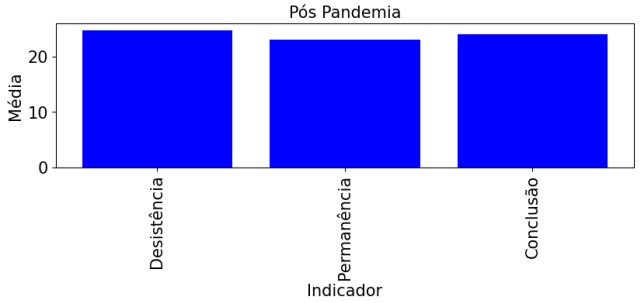

Sexo - Permanência
Desistência
M
60.0%
40.0%
F
Permanência
M
45.2%
54.8%
F
Conclusão
M
46.4%
53.6%
F
Feminino
Masculino


Sexo - Conclusão
Desistência
M
46.9%
53.1%
F
Permanência
M
45.2%
54.8%
F
Feminino
Masculino
Conclusão
M
100.0%

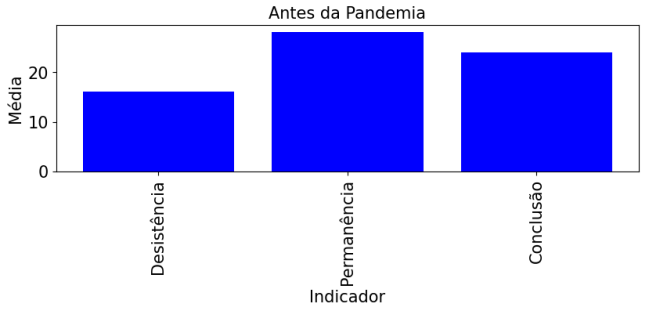
Idade média X indicador

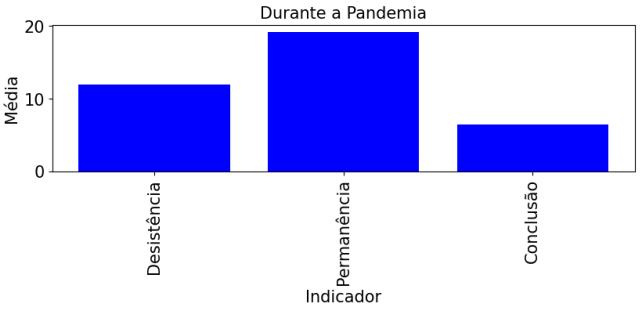


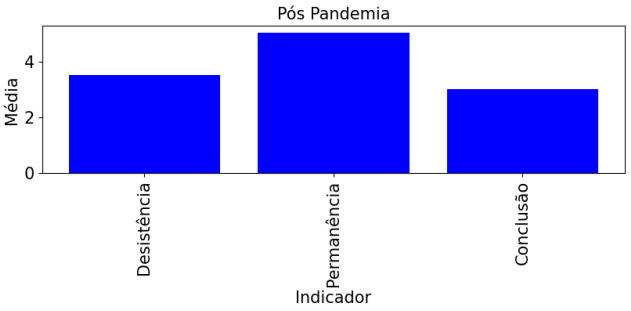


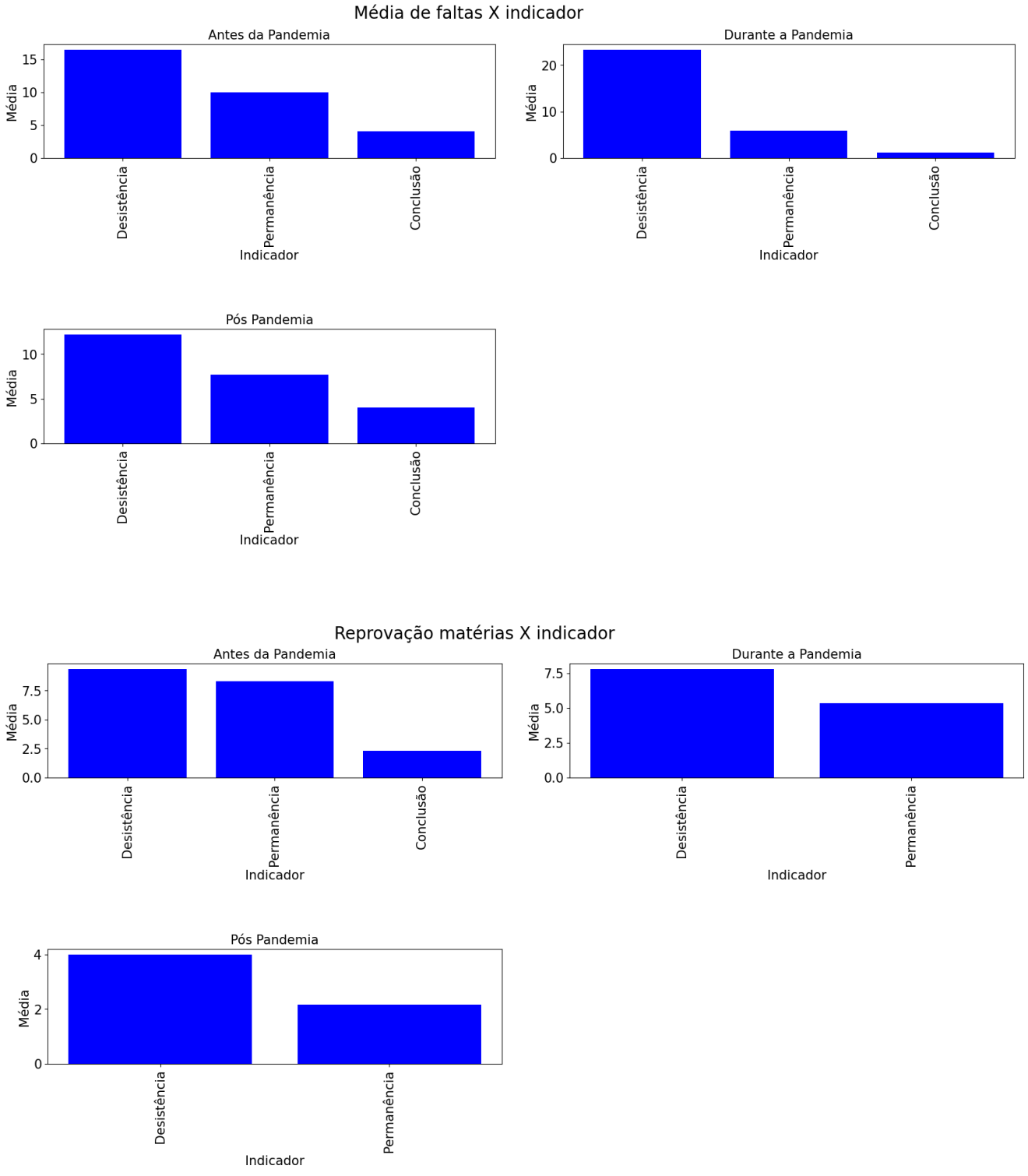


Média de disciplinas cursadas X indicador









Fonte: Elaborado pelos autores, 2023.