UNIVERSIDADE NOVE DE JULHO – UNINOVE

PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA E GESTÃO DO CONHECIMENTO

AMANDA FERREIRA DE MOURA

MINERAÇÃO DE DADOS EDUCACIONAIS PARA APOIO À GESTÃO ACADÊMICA NA FORMULAÇÃO DE PROGNÓSTICO DE PERFIL DE ALUNO INGRESSANTE EM CURSOS SUPERIORES

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Informática e Gestão do Conhecimento da Universidade Nove de Julho - UNINOVE, como requisito parcial para obtenção do título de Mestre.

Prof. Orientador: Dr. Marcos Antonio Gaspar

São Paulo

2023

AMANDA FERREIRA DE MOURA

MINERAÇÃO DE DADOS EDUCACIONAIS PARA APOIO À GESTÃO ACADÊMICA NA FORMULAÇÃO DE PROGNÓSTICO DE PERFIL DE ALUNO INGRESSANTE EM CURSOS SUPERIORES

Pesquisa de dissertação apresentada ao Programa de Pós-Graduação em Informática e Gestão do Conhecimento da Universidade Nove de Julho – UNINOVE, como requisito parcial para a obtenção do grau de Mestre em Informática e Gestão do Conhecimento.

Prof. Orientador: Dr. Marcos Antonio Gaspar

São Paulo

2023

Moura, Amanda Ferreira de.
Mineração de dados educacionais para apoio à gestão acadêmica na formulação de prognóstico de perfil de aluno ingressante em cursos superiores. / Amanda Ferreira de Moura. 2023.
175 f.
Dissertação (Mestrado) - Universidade Nove de Julho - UNINOVE, São Paulo, 2023.
Orientador (a): Prof. Dr. Marcos Antonio Gaspar.
1. Inteligência artificial. 2. Mineração de dados. 3. Mineração de dados educacionais. 4. Gestão acadêmica. 5. Ensino superior.
I. Gaspar, Marcos Antonio. II. Título.
CDU 004

UNINOVE
Universidade Nove de Julho

PARECER – EXAME DE DEFESA

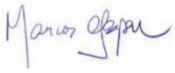
Parecer da Comissão Examinadora designada para o exame de defesa do Programa de Pós-Graduação em Informática e Gestão do Conhecimento a qual se submeteu a aluna Amanda Ferreira de Moura.

Tendo examinado o trabalho apresentado para obtenção do título de "Mestre em Informática e Gestão do Conhecimento", com Dissertação intitulada "MINERAÇÃO DE DADOS EDUCACIONAIS PARA APOIO À GESTÃO ACADÊMICA NA FORMULAÇÃO DE PROGNÓSTICO DE PERFIL DE ALUNO INGRESSANTE EM CURSOS SUPERIORES", a Comissão Examinadora considerou o trabalho:

(X) Aprovado
( ) Aprovado condicionalmente
( ) Reprovado com direito a novo exame
( ) Reprovado

EXAMINADORES

Prof. Dr. Marcos Antonio Gaspar - Uninove (Orientador)



Prof. Dr. Ivanir Costa - Uninove (Membro Interno)

Prof. Dr. Fabio Kazuo Ohashi - Universidade Presbiteriana Mackenzie (Membro Externo)

São Paulo, 28 de junho de 2023.

Dedico esta dissertação a Deus por me dar forças para enfrentar este grande desafio e acreditar mim, mesmo quando eu não acreditava e por manter-se presente em cada etapa da minha vida.

AGRADECIMENTOS

Desejo expressar mais uma vez meu agradecimento a Deus, por me ajudar até aqui, e me sustentar durante os momentos difíceis ao longo deste caminho.

Agradeço à minha amada família, que mesmo diante de desafios, sempre me proporcionou um apoio inabalável, permitindo que eu alcançasse com êxito esta jornada acadêmica.

Aos meus preciosos amigos, quero transmitir minha profunda gratidão por seu contínuo apoio e incentivo ao longo de todo o trajeto. Suas palavras de encorajamento e presença constante foram um farol de esperança nos momentos mais exigentes.

Gostaria de expressar minha profunda gratidão com imenso carinho ao meu primeiro orientador, o Prof. Dr. Domingos Márcio Rodrigues Napolitano, por ter acreditado em meu projeto e me acolhido como orientanda no início desta jornada. Sua orientação inicial foi fundamental, proporcionando a base essencial para o meu crescimento profissional.

Manifesto com profunda estima minha gratidão ao meu atual orientador, o Prof. Dr. Marcos Antonio Gaspar, cuja orientação sábia e paciente foram pilares fundamentais para minha conquista acadêmica. Sua confiança e dedicação a este projeto foram inestimáveis, impulsionando-me a superar cada obstáculo com comprometimento excepcional com a excelência.

Aos ilustres membros da banca, Prof. Dr. Ivanir Costa e Prof. Dr. Fabio Kazuo Ohashi, expresso minha gratidão pela disponibilidade e pela generosidade habitual, além das valiosas contribuições que enriqueceram sobremaneira este trabalho.

Por último, mas não menos importante, minha profunda gratidão à Universidade Nove de Julho, cuja estrutura e investimento em pesquisa abriram portas para a realização deste sonho em minha vida. Seu apoio institucional foi essencial para que eu pudesse trilhar este caminho de aprendizado e crescimento.

"Lembre-se: olhe para as estrelas, e não para os seus pés."

(Stephen Hawking)

RESUMO

As instituições de ensino são organizações que geram grande volume de dados em seus procedimentos rotineiros. A gestão desses dados educacionais é uma tarefa importante para propiciar acompanhamento da performance do aluno, bem como proporcionar prognósticos para ações a serem tomadas pelo gestor. A mineração de dados educacionais é uma vertente da mineração de dados que se propõe a extrair conhecimento dos dados gerados nas instituições de ensino. O objetivo desta pesquisa foi desenvolver uma solução automatizada de mineração de dados educacionais para apoio à gestão acadêmica na formulação de prognóstico de perfil de aluno ingressante em cursos superiores. Para atingir tal objetivo foi realizada pesquisa experimental aplicada de natureza quantitativa. Assim sendo, foram realizados experimentos computacionais com a aplicação de técnicas inteligentes de mineração de dados voltadas ao agrupamento (clusterização) de dados educacionais de uma base de dados de alunos de cursos de pós-graduação. Esta solução foi desenvolvida em quatro etapas, sendo indicadas em cada fase as respectivas ferramentas, bases de dados, operações e ações necessárias. Os resultados produzidos pelos experimentos realizados na aplicação da mineração de dados comprovaram a eficiência da solução concebida, ou seja, a solução de mineração de dados educacionais desenvolvida nesta pesquisa tem capacidade de estabelecer o prognóstico do perfil de aluno ingressante mais aderente e alinhado aos cursos disponibilizados pela instituição de ensino superior. Cabe ressaltar que a solução desenvolvida se apoia em ferramentas gratuitas ou de baixo custo, o que a torna acessível às instituições que queiram implementar esta aplicação, bem como criar soluções baseadas em mineração de dados educacionais.

Palavras-chave: Inteligência artificial. Mineração de dados. Mineração de dados educacionais. Gestão acadêmica. Ensino superior.

ABSTRACT

Educational institutions are organizations that generate large amounts of data in their routine procedures. The management of these educational data is an important task to provide student performance monitoring, as well as providing prognoses for actions to be taken by the manager. Educational data mining is a branch of data mining that aims to extract knowledge from data generated in educational institutions. The objective of this research was to develop an automated educational data mining solution to support academic management in the formulation of a prognostic profile for students entering higher education courses. To achieve this objective, applied experimental research of a quantitative nature was carried out. Therefore, computational experiments were performed with the application of intelligent data mining techniques aimed at grouping (clustering) educational data from a database of postgraduate students. This solution was developed in four stages, with the respective tools, databases, operations and necessary actions being indicated in each stage. The results produced by the experiments carried out in the application of data mining proved the efficiency of the conceived solution, that is, the educational data mining solution developed in this research has the capacity to establish the prognosis of the freshman student profile that is more adherent and aligned with the available courses by the higher education institution. It should be noted that the developed solution is based on free or low-cost tools, which makes it accessible to institutions that want to implement this application, as well as create solutions based on educational data mining.

Keywords: Artificial intelligence. Data mining. Educational data mining. Academic management. University education.

LISTA DE ABREVIAÇÕES

EEMN	Escolaridade Ensino Médio Normal
EEMT	Escolaridade Ensino Médio Técnico
EEME	Escolaridade Ensino Médio EJA
EGB	Escolaridade Graduação Bacharel
EGT	Escolaridade Graduação Tecnólogo
EPG	Escolaridade Pós-Graduação
DG1	Nota da disciplina Gerencial 1
DG2	Nota da disciplina Gerencial 2
DG3	Nota da disciplina Gerencial 3
DG4	Nota da disciplina Gerencial 4
DG5	Nota da disciplina Gerencial 5
DG6	Nota da disciplina Gerencial 6
MFG	Média final das disciplinas gerenciais
DT1	Nota da disciplina Técnica 1
DT2	Nota da disciplina Técnica 2
DT3	Nota da disciplina Técnica 3
DT4	Nota da disciplina Técnica 4
MFT	Média final das disciplinas técnicas
MF	Média final total das disciplinas (gerenciais e técnicas)
SIT	Situação (aprovado, reprovado ou trancou o curso)
CC	Ciência da Computação
SI	Sistemas de Informação
TADS	Tecnologia em Análise e Desenvolvimento de Sistemas
TGTI	Tecnologia em Gestão da Tecnologia da Informação
TRC	Tecnologia em Redes de Computadores
TSEG	Tecnologia em Segurança da Informação
TSIN	Tecnologia em Sistemas para Internet

LISTA DE FIGURAS

LISTA DE QUADROS

SUMÁRIO

1. INTRODUÇÃO

1.1 Contextualização

Em meio ao processo de constantes mudanças no ensino e na busca de conhecimento na sociedade, o reconhecimento da importância da educação superior tem se evidenciado a partir da metade do século XX, uma vez que ocorreram importantes mudanças e transformações nos sistemas nacionais de educação no Brasil. Nas últimas décadas, o significativo crescimento da educação superior demonstrou que uma maior qualificação de recursos humanos aumenta de forma considerável as taxas de retorno no setor trabalhista em termos de rendimento e empregabilidade (BROCH; BRESCHILIARE; BARBOSA-RINALDI*,* 2020). Desta forma, os governos de diferentes instâncias e regiões passaram a conceber a educação superior como um fator vital para o desenvolvimento e competitividade econômica entre os países (OLIVEIRA, 2014; NEVES, 2018). Além do crescimento no aspecto econômico, também ficou evidente a melhora no aspecto social, uma vez que o acesso à educação superior se configurou como um importante mecanismo para a redução da desigualdade social e elevação de oportunidades e para a promoção da mobilidade social (VÉRÉTOUT, 2012; NEVES, 2018).

Segundo o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira - INEP - (2020), ao final de 2020 as instituições de educação superior (IES) alcançaram o número de 2.457 organizações ativas, com crescimento desde 2011 na ordem de 3,9%. Porém, concluindo o ano de 2020, comparando-se a 2019, houve uma diminuição de 151 instituições de educação superior (IES), o que representou um decréscimo de 5,8%, provocado principalmente pelos impactos da pandemia de COVID-19 a partir deste ano. Dietrich, *et al.,* (2020) evidencia a grande mudança ocorrida no ensino neste período, uma vez que 1,7 bilhão de alunos em todo mundo sofreram uma interrupção em seus estudos presenciais, o que provocou uma revolução para educadores e instituições de ensino superior que, devido à urgência da situação, foram obrigadas a adequar suas aulas para novas modalidades de ensino.

Contudo, mesmo com este contexto pandêmico, a oferta por cursos de graduação cresceu para acompanhar a evolução das instituições, com o total de 37,9% de aumento no período de 2011 a 2020, sendo uma expressiva maioria em cursos à distância (INEP, 2020). Tal fato concebeu um novo olhar para educação à distância no Brasil, que têm intensificado seus esforços na criação e adaptação de conteúdos educacionais de forma dinâmica e variada. Em geral, acredita-se que este tipo de ensino, além de estar em crescimento, vem colaborando com a democratização do saber. Isto porque esta modalidade de oferta de ensino alcança um número maior de sujeitos simultaneamente. Tal fenômeno proporciona assim oportunidades de acesso para indivíduos situados em locais distantes da sede da instituição de ensino superior que oferece o serviço de educação (KUBRUSLY, 2021).

Por conta do crescimento do ensino superior evidenciado até aqui, constatou-se também como consequência o aumento no volume de dados acadêmicos gerados durante a jornada do aluno no curso. Nesse sentido, atualmente estão disponíveis tecnologias voltadas ao processamento inteligente de grandes massas de dados (*big data*), que trazem possibilidade de análise de um expressivo conjunto de dados rapidamente e com segurança. Assim, os métodos e técnicas de Inteligência Artificial provêm o tratamento da modelagem e predição a partir de centenas de variáveis presentes em grandes volumes de dados.

Baker e Smith (2019, p.10) entendem que a Inteligência Artificial seja formada por "computadores que realizam tarefas cognitivas, geralmente associadas a mentes humanas, particularmente aprendizagem e resolução de problemas". O setor educacional tem incorporado estas tecnologias para o tratamento de dados educacionais em virtude das variadas formas de abordagens e aplicações no ensino superior, visando assim criar informações úteis para um melhor planejamento letivo e gestão acadêmica (SILVA; FONSECA, 2017).

A análise de dados educacionais tem sido explorada em diferentes pesquisas em prol de um maior aproveitamento dos dados gerados a cada ano letivo. Na área de Informática em Educação cresce a quantidade de pesquisas que promovem o desenvolvimento de ferramentas exploratórias para gerir as ofertas de cursos, assim como entender melhor os discentes e os cenários em que eles aprendem (BRITO, 2015). Como pesquisa, a investigação de dados educacionais se desenvolve em

temas como *Educational Data Mining* - EDM (Mineração de Dados Educacionais), *Learning Analytics* - LA (Análise de Aprendizagem) e *Academic Analytics* - AA (Análise Acadêmica); sendo estas frentes voltadas à tomada de decisão no âmbito da gestão educacional (ARANA, 2018).

A Inteligência Artificial aplicada ao processo de ensino-aprendizagem para fins de gestão acadêmica obteve maior atenção a partir de 2018. O relatório Educause Horizon Report 2018 (BECKER, 2018) sinalizava naquele momento um crescimento de 43% no período de 2018-2022. Mas já em 2019, este mesmo estudo (ALEXANDER, 2019) constatou que o crescimento seria maior que o previsto, com intensos investimentos aplicados por empresas privadas, a exemplo do Google. O estudo evidenciou ainda ser inevitável a necessidade da aplicação da Inteligência Artificial ao processo de ensino-aprendizagem voltado à gestão acadêmica (CONTACT NORTH, 2018; ZAWACKI-RICHTER, *et al.*, 2019).

A Mineração de Dados Educacionais (EDM) tem a finalidade de descobrir informações que possam auxiliar a proposta educacional, trazendo melhor aproveitamento do processo de ensino-aprendizagem, bem como auxiliar na previsão do desempenho do aluno, além de influenciar a aprendizagem do aluno (BAKER *et al.*, 2011; DO NASCIMENTO, 2018).

Há necessidade de gerir dados educacionais/acadêmicos, tanto para diagnóstico dos resultados alcançados, quanto para prognóstico de decisões a serem tomadas por gestores nas instituições de ensino superior. Para tanto, as instituições de ensino superior enredaram esforços na criação de sistemas de gerenciamento de banco de dados educacionais, que possibilitam administrar grande volume e fluxo de dados, além do uso de mineração de dados para a extração de informações relevantes para o processo de tomada de decisão quanto a oferta de novos cursos. Siemens e Baker (2012) e Du, Hung e Shelton (2019) realizaram estudos sobre como extrair informações valiosas para automatizar o processo de aprendizagem ou implementações de intervenção em cursos. Este tipo de automação inteligente da Mineração de Dados Educacional (EDM) está amparada por técnicas de Inteligência Artificial.

Desta forma, conforme argumentos de Liñán e Pérez (2015) e Dutt, Ismail e Herawan (2017), a Mineração de Dados Educacionais (EDM) se configura num novo

e amplo campo de pesquisa. Anoopkumar e Rahman (2016) apontam que atualmente a EDM desempenha importante papel na descoberta de padrões de conhecimento sobre fenômenos educacionais no contexto da gestão acadêmica. Em especial, a mineração de dados educacionais tem sido usada para prever uma variedade de resultados educacionais, tais como desempenho do aluno e aderência do curso ao perfil do aluno (TANG; XING; PEI, 2019).

1.2 Problema de Pesquisa

A explosão do volume de dados gerados e coletados durante o processo de ensino-aprendizagem de alunos sinaliza para a necessidade de instrumentalizar o tratamento e análise destes dados, de forma a possibilitar o aperfeiçoamento da tomada de decisões por parte da gestão acadêmica de instituições de ensino superior. Na visão de Shelton, *et al.,* (2017) e Dimic, *et al.,* (2019), a mineração de dados educacionais pode oferecer suporte para a gestão acadêmica obter uma visão holística do progresso de aprendizagem do aluno, além de viabilizar acompanhamento personalizado do aluno, baseado em evidências ou em dados, permitindo assim intervenções ou recomendações nos serviços oferecidos.

Considerando-se o contexto até então apresentado, a presente dissertação busca responder à seguinte questão de pesquisa: *como desenvolver e validar uma solução automatizada de mineração de dados educacionais para apoio à gestão acadêmica na formulação de prognóstico de perfil de aluno ingressante em cursos superiores?*

1.3 Objetivos de Pesquisa

Esta dissertação tem como objetivo desenvolver uma solução automatizada de mineração de dados educacionais para apoio à gestão acadêmica na formulação de prognóstico de perfil de aluno ingressante em cursos superiores.

Em complemento ao objetivo geral, os seguintes objetivos específicos são indicados:

1) Identificar na literatura acadêmica sustentação teórica da temática abordada nesta pesquisa: inteligência artificial, mineração de dados, mineração de dados educacionais e gestão acadêmica;
2) Desenvolver uma solução automatizada de mineração de dados educacionais para suporte à gestão acadêmica de cursos superiores;
3) Validar uma solução automatizada de mineração de dados educacionais para suporte à gestão acadêmica de cursos superiores;
4) Elaborar um guia de orientação dos requisitos para implementação da solução automatizada de mineração de dados educacionais para suporte à gestão acadêmica de cursos superiores.

1.4 Justificativa

Atualmente as instituições de ensino superior enfrentam exigências cada vez maiores na qualidade do ensino oferecido aos discentes. Diante da necessidade de melhorias constantes na qualidade, é importante buscar estratégias para uma gestão acadêmica eficiente. Para tanto, a gestão acadêmica com base na mineração de dados educacionais (EDM), apoiada por métodos, técnicas e ferramentas de inteligência artificial pode proporcionar melhorias na oferta de cursos e aproveitamento dos serviços prestados pelos alunos (MAJERNÍK, *et al.*, 2022).

Por outro lado, a Mineração de Dados Educacionais (EDM) é um campo de estudos interdisciplinar ainda recente, com muito espaço para se desenvolver. Atualmente as instituições de ensino coletam e arquivam grandes quantidades de dados, como registro de alunos, frequência e resultados de exames. A mineração desses dados pode auxiliar as instituições de ensino a entender o comportamento e os interesses dos alunos, bem como seu alinhamento ao curso em que se encontram matriculados. Isto é possível pela extração de informações úteis a partir da mineração dos dados educacionais disponíveis (GUPTA, 2020).

Segundo Kumar e Sharma (2017, p. 2), a mineração de dados educacionais (EDM) refere-se aos “métodos e ferramentas computadorizadas para detectar e extrair automaticamente padrões e informações significativas de grandes coleções de dados de ambientes educacionais”. Vários algoritmos de aprendizado de máquina e

ferramentas de pesquisa são empregados na mineração de dados educacionais voltada à análise e previsão de diferentes tipos de dados educacionais.

A mineração de dados educacionais (EDM) é um campo de estudos emergente e novas ferramentas e técnicas têm sido desenvolvidas continuamente. Mahajan e Manipal (2020) apontam as principais ferramentas de mineração de dados educacionais aplicadas na atualidade: RapidMiner, Knime, Orange, SPSS, KEEL, The EDM Workbench, Spark MLLib, D3js e PSLC DataShop.

É importante reforçar os potenciais contribuições desta dissertação para a evolução das pesquisas sobre a aplicação de mineração de dados educacionais. Em adição, também destaca-se a criação e validação de uma solução automatizada de mineração de dados educacionais para suporte à gestão acadêmica de cursos superiores. Tal solução poderá ser utilizada por gestores de instituições de ensino superior para diagnóstico e prognóstico de alunos dos cursos oferecidos, visando assim alterar ou adaptar as características do curso, bem como entender qual o perfil mais adequado de aluno para cada curso. Desta forma, vislumbra-se ser possível aos gestores acadêmicos proporem cursos mais adequados conforme o perfil de aluno mais pertinente.

2. REFERENCIAL TEÓRICO

Neste capítulo, serão apresentados os principais tópicos abordados nesta pesquisa, juntamente com os principais autores dispostos na literatura.

2.1 Inteligência Artificial Aplicada à Gestão Acadêmica

Para Schafer e Kaufman (2018), a Inteligência Artificial (IA) diz respeito a um campo de conhecimento associado à linguagem e à inteligência, ao raciocínio, à aprendizagem e à resolução de problemas. O desenvolvimento da IA ocorre em estágios junto à expectativa de resultados com sua prática, que podem variar dependendo de suas aplicações.

Conforme indicado por Sandhya e Prasath (2019) e Kaviya e Premlatha (2019), a IA é um subcampo da ciência da computação direcionado à resolução de problemas de forma a executar tarefas humanas. A título de exemplo, é possível indicar que o ser humano possui a capacidade de reconhecer formas e objetos e assim entender padrões. Na IA esta ação vem por combinações de dados e objetos classificados apresentados, criando-se assim uma ‘memória’ de aprendizado.

Segundo Holmes e Wayne (2020), a IA é a força tecnológica motriz da primeira metade deste século e transformará praticamente todos os setores da sociedade. Empresas e governos em todo o mundo estão realizando significativos investimentos em uma ampla gama de implementações com uso de IA de diversos setores, desde instituições de ensino até *startups* que estão em crescimento na ordem de bilhões de dólares.

Nas instituições de ensino superior, a geração de dados dos alunos continua a aumentar a cada ano. Assim, há a necessidade do uso de IA para analisar os dados de forma a utilizá-los para melhorar os processos de aprendizagem e gestão dessas instituições. Com a ajuda da mineração de dados educacionais, as instituições de ensino superior podem ser dotadas de meios eficazes para melhorar a eficácia da instituição e o processo de aprendizagem dos alunos (HUEBNER, 2013; BAKER, 2014; HOLMES; WAYNE, 2020).

Especificamente quanto à aplicação de IA no setor de educação, muitos trabalhos de pesquisa foram realizados neste campo por diversos pesquisadores que fizeram uso de diferentes técnicas de mineração para a extração de dados em bancos

de dados educacionais. Alsuwaiket, *et al.,* (2020) indicam as principais técnicas de mineração aplicadas por pesquisadores na área de educação, a saber: Árvores de Decisão, Árvores de Regressão, Cadeias de Markov, Regras de Associação, Regressão Linear, Padrões Sequenciais, Análise de Correlação, Redes Bayesianas, Redes Neurais Artificiais, Classificação, Clustering, Mineração de Sequência Diferencial, Lógica Fuzzy e Algoritmos Genéticos.

Nos últimos anos as técnicas de IA e Aprendizado de Máquina (ML) têm se tornado mais populares entre os pesquisadores para aplicação na mineração de dados educacionais. A razão por trás do incremento dessas aplicações se dá em razão de haver um amplo escopo para melhorar os resultados usando tais técnicas de IA na mineração de dados educacionais (CHU, *et al.*, 2019).

A título de exemplos de estudos com aplicação de IA na gestão acadêmica, Alsuwaiket, *et al.,* (2020) desenvolveram um modelo preditivo com classificadores *Random Forest* e *Naive Bayes* para resolver o problema da lacuna entre o curso e a avaliação baseada em exames aplicados aos alunos. Os autores observaram que este índice ajuda a aumentar a precisão da previsão da média obtida pelos alunos no segundo ano do curso, com base na média obtida no primeiro ano. Chu, *et al.,* (2019) aplicaram técnicas de mineração de dados para aplicação nos dados coletados no Departamento de Engenharia Civil da Ho Chi Minh City University of Transport (Vietnã) no período 2013-16. Tal aplicação voltava-se a apoiar os alunos na escolha dos cursos. Verificou-se que os resultados dos experimentos realizados podem trazer consequências positivas para a escolha do curso.

Outras técnicas de IA também são aplicadas para trazer melhorias no processo de gestão acadêmica. Rogers (2019) aplicou a técnica difusa (Regressão Logística e *Real Fuzzy Linear*) para classificar os alunos com base no desempenho individual. Foram reunidos dados de 172 alunos, de quatro escolas primárias em Blackbelt of Alabama e Mississippi (EUA). As respostas da pesquisa junto aos alunos em relação à avaliação dos professores foram usadas como dados. Como resultado, o modelo aplicado foi capaz de estrutura uma classificação com sucesso de até 90%.

Com o uso de Árvore de Decisão (*Decision Tree*), J48 e Árvores Aleatórias (*Random Trees*), Moscoso-Zea *et al.* (2019) buscaram encontrar uma solução para prever a taxa de graduação em alunos de uma universidade privada. Os dados dos

alunos foram coletados no Departamento de Ciência da Computação. As árvores aleatórias forneceram resultados precisos, tendo sido a possibilidade de interpretação dos resultados com melhor desempenho a que utilizou o algoritmo J48. O estudo conduzido por Adekitan e Salau (2019) visando determinar o impacto do desempenho dos alunos em seus resultados baseou-se na avaliação dos primeiros três anos acadêmicos de 1.841 alunos de uma universidade. Foram aplicadas as ferramentas NN, Floresta Aleatória, Árvore de Decisão, *Nave Bayes* e Regressão Logística. Obteve-se resultados positivos com acertos entre de 85,89% até 89,15% de precisão.

Para ajudar as instituições na seleção e aplicação de melhores práticas de ensino e aprendizagem, Rahman, *et al*., (2018) coletaram dados por meio de questionário composto por 38 questões relacionadas ao ensino e aprendizagem em uma instituição de ensino. Como resultado, observou-se que mais aspectos da mineração de dados educacionais foram abordados aplicando-se estes algoritmos, quando comparados a outras ações.

A Inteligência Artificial (IA) desempenha um papel significativo na gestão acadêmica, oferecendo soluções avançadas para melhorar a eficiência e a eficácia das instituições de ensino. A IA pode ser aplicada a diferentes áreas da gestão acadêmica, desde a coleta e análise de dados até a automação de processos administrativos. A título de exemplo indica-se que algoritmos de aprendizado de máquina podem ser utilizados para identificar padrões em grandes conjuntos de dados educacionais, permitindo uma compreensão mais profunda do desempenho dos alunos, a detecção de problemas potenciais e a personalização do ensino (CROMPTON; HELEN; BURKE, 2023).

Além disso, a IA pode desempenhar um papel importante na automação de tarefas administrativas. Chatbots e assistentes virtuais podem ser usados para responder a perguntas comuns dos alunos, orientar em procedimentos acadêmicos e fornecer suporte 24 horas por dia, 7 dias por semana. Tais aplicações reduzem a carga de trabalho dos funcionários e melhoram a experiência do aluno (ALYAHYAN; EYMAN; DUSTEGOR, 2020). Segundo Pedro *et al.* (2019), a IA também pode ser aplicada no planejamento acadêmico e na alocação de recursos em instituições de ensino. Isto porque os algoritmos de otimização podem ajudar na programação de horários de aulas, alocação de salas, distribuição de docentes e gestão de recursos

financeiros. Tais aplicações ajudam a melhorar a eficiência dos processos e a maximizar o uso dos recursos disponíveis, proporcionando assim uma gestão acadêmica mais eficaz.

Por fim, depreende-se que a IA oferece uma variedade de oportunidades para melhorar a gestão acadêmica em diferentes tipos de instituições de ensino, desde a análise de dados até a automação de processos e otimização de recursos. A partir da aplicação adequada da IA, as instituições de ensino podem aprimorar a tomada de decisão, a eficiência operacional e a experiência do aluno, contribuindo assim para impulsionar a qualidade e a inovação na educação.

2.2 Mineração de Dados Aplicada à Gestão Acadêmica

A aplicação de técnicas de mineração de dados tem aumentado na área educacional, pois pode ajudar a promover a melhoria do sistema educacional ao contribuir para a evolução dos alunos por meio da indicação de desempenho, bem como por meio da elaboração de previsões e prognósticos. A Mineração de Dados Educacional (*Educational Data Mining* - EDM) é um campo recente de caráter interdisciplinar que auxilia profissionais e instituições a lidar com os dados relacionados à perspectiva educacional. Atualmente, as instituições de ensino coletam dados estruturando-os em bases de dados com elevado volume de registros de alunos, tais como frequência e resultados de provas durante sua jornada acadêmica. A mineração desses dados ajuda as instituições de ensino a entenderem o comportamento e os interesses dos alunos, extraindo informações úteis desses enormes volumes de dados disponíveis (GUPTA, *et al.,* 2020).

Diferentes técnicas de mineração de dados estão sendo aplicadas para a mineração de dados no campo educacional. Atualmente, as técnicas de Inteligência Artificial e Aprendizado de Máquina são mais populares entre os pesquisadores para com aplicações voltadas à criação de informações a partir dos bancos de dados educacionais, pois fornecem resultados mais confiáveis em comparação com outras técnicas (SILVA; FONSECA, 2017).

Os diferentes métodos de mineração de dados utilizados são calcados em técnicas estatísticas baseadas em probabilidade e matemática, tais como *Naive*

Bayes, Regressão Linear e Logística, técnicas de aprendizado de máquina como aprendizado supervisionado: *Support Vector Machine* (SVM) e Árvore de Decisão (DT) e aprendizado não supervisionado: *Clustering* em *Fuzzy Logic* e *Artificial Neural Network* (ANN) e *Association Rule Mining*, dentre outras (MOHAMAD; TASIR, 2013; MANJARRES, *et al.,* 2018).

A aplicação de mineração de dados educacionais no contexto de instituições de ensino pode auxiliar na melhoria do processo de ensino e aprendizagem, por meio da extração de informações úteis dos dados educacionais (ABU TAIR; EL-HALEES, 2012).

A mineração de dados é aplicada em diferentes áreas, empregando grandes fluxos de dados e algoritmos específicos voltados à extração e análise de dados. Hung, *et al.,* (2020) aplicaram a mineração de dados para explorar os comportamentos de aprendizagem em dados gerados por alunos em um curso de ensino híbrido. Os dados experimentais foram coletados de duas turmas de cursos relacionados à programação em Python para alunos do primeiro ano em uma universidade de Taiwan. A aplicação de modelo baseado em *Random Forest* possibilitou identificar, durante o semestre, os alunos em alto risco de reprovação. Isto foi possível a partir da precisão pelos dados gerados no ambiente educacional em questão. Além disso, foram utilizados também aprendizado de máquina e algoritmos de aprendizado baseados em simetria para explorar os comportamentos de aprendizado dos alunos.

A tecnologia e a inovação capacitam as instituições de ensino superior (IES) a usarem diferentes tipos de sistemas de aprendizagem, a exemplo do *videolearning*. Em pesquisas foram utilizadas técnicas de mineração de dados de modo a melhorar o desempenho e aproveitamento dos alunos. Estudo conduzido por Hasan, *et al.,* (2020) foi realizado com 772 exemplos de alunos matriculados nos módulos de e-commerce e tecnologias de e-commerce de uma IES. O estudo teve como objetivo prever o desempenho geral do aluno ao final do semestre, usando para tanto a análise de aprendizado de vídeo e técnicas de mineração de dados. Os dados do sistema de informação do aluno, sistema de gerenciamento de aprendizagem e aplicativos móveis foram analisados aplicando-se oito algoritmos de classificação diferentes. Também foram aplicadas técnicas de transformação e pré-processamento de dados visando reduzir as características dos dados originais. Como resultado, a técnica

Random Forest indicou os alunos bem-sucedidos ao final do semestre, com uma precisão de 88,3%.

Para Khan e Ghosh (2021), a modelagem de desempenho de aluno é um dos tópicos de pesquisa desafiadores e largamente aplicados em estudos de mineração de dados educacionais (EDM). Diversos fatores influenciam o desempenho de formas não lineares, o que contribui para tornar este campo mais atraente para os pesquisadores. Neste sentido, os autores apresentaram resultados de revisão sistemática de literatura sobre estudos de EDM acerca do desempenho do aluno na aprendizagem em sala de aula. Para tanto, tal estudo abarcou a identificação dos preditores, métodos utilizados para tal identificação, tempo e objetivo da predição. Com uma base de 140 estudos nesta área, os resultados indicaram que os pesquisadores alcançaram uma eficiência de previsão significativa durante o período do curso. No entanto, a previsão de desempenho antes do início do curso necessita atenção especial.

Segundo Alqasemi, *et al.,* (2021), nos últimos anos a Mineração de Dados Educacionais (EDM) se tornou um novo campo inserido na mineração de dados. A EDM tem sido empregada para extrair novas informações e conhecimentos intrínsecos ao processo educacional, razão pela qual a EDM se tornou um tema proeminente no campo da informática educacional. Em seu estudo, os autores aplicaram a análise de agrupamento nas estatísticas educacionais de regiões do Iêmen, a fim de viabilizar um processo de mineração usando algoritmo hierárquico. A análise de agrupamento retratou um conhecimento latente sob os dados educacionais, que foi ilustrado por um dendrograma ou diagrama hierárquico. Os resultados desse estudo apresentaram relações promissoras entre as regiões do Iêmen, que ajudariam os gestores responsáveis pela tomada de decisão a entender a natureza das variáveis educacionais, que estão distribuídas pelo país.

2.3 Mineração de Dados Educacionais (EDM)

As técnicas de mineração de dados são cada vez mais aplicáveis a questões do setor educacional. Como em muitos outros setores, o ensino superior tem se atentado para o impacto potencial dessas técnicas no processo e nos resultados da

aprendizagem, a fim de avançar em termos de ferramentas tecnológicas em prol da melhoria nos processos de ensino-aprendizagem da educação formal. Nesse sentido, a aplicação de diferentes técnicas de mineração de dados pode ser vista, como um potencial alicerce para uma mudança sistêmica. Além disso, tal aplicação pode impactar positivamente se for considerada um instrumento de auxílio às instituições de ensino superior para encontrarem soluções para seus problemas específicos (VAN BARNEVELD, *et al.*, 2012; ALDOWAH, *et al*., 2019).

Assim, os resultados de aplicações de mineração de dados na educação podem fornecer importante suporte ao processo de tomada de decisão em instituições de ensino (PEÑA AYALA, 2014). Isto porque a mineração de dados educacionais (EDM) e a análise de aprendizagem são duas áreas específicas que combinadas ajudam a representar o uso e a aplicação da mineração de dados no ensino superior e outros ambientes educacionais. Assim, a combinação de mineração de dados educacionais (EDM) e análise de aprendizagem estabelece um ecossistema que pode coletar, processar, relatar e trabalhar continuamente em dados digitais para melhorar o processo educacional. Tal combinação pode fornecer aos formuladores de políticas educacionais modelos baseados em dados essenciais para apoiar seus objetivos de aumentar a eficiência da aplicação do ensino (BARRETO, 2020).

Segundo Hernández-Blanco, *et al.,* (2019), a Mineração de Dados Educacionais (EDM) está preocupada em desenvolver, pesquisar e aplicar aprendizado de máquina, mineração de dados e métodos estatísticos para detectar padrões em grandes coleções de dados educacionais que, de outra forma, seriam impossíveis de analisar. Nesse sentido, Bakhshinategh, *et al.,* (2018) afirma existirem diferentes métodos e aplicações em EDM que podem seguir tanto objetivos de pesquisa aplicada, como a melhoria da qualidade da aprendizagem; quanto objetivos de pesquisa pura, mais voltados a melhorar a compreensão do processo de aprendizagem.

Na visão de Aldowah *et al.* (2019), as pesquisas realizadas demonstram a importância da mineração de dados educacionais nas instituições de ensino. O conhecimento adquirido pelo uso de técnicas de mineração de dados pode ser aplicado para tomar decisões bem-sucedidas e eficazes que irão melhorar e fazer progredir o desempenho do aluno na educação. Em estudo conduzido por Jalota e

Agrawal (2019) foi utilizado um conjunto de dados educacionais contendo 163 instâncias e dezesseis atributos. Foram testados cinco classificadores com o uso da ferramenta *Weka* para comparações feitas com base na precisão entre esses classificadores, considerando-se diferentes medidas para determinar qual o melhor classificador. Os resultados dos experimentos mostraram que o *Multilayer Perceptron* tem o melhor desempenho entre outros classificadores.

Em estudo de Sunita e Jawandhiya (2022), comparou-se diferentes ferramentas de mineração de dados gratuitas de código aberto e proprietárias. Essas ferramentas foram comparadas com base no sistema operacional, tipo, linguagem em que a ferramenta foi escrita e usos, dentre outras características. As ferramentas consideradas no estudo foram Keel, KNIME, RapidMiner, Weka, Tanagra e Orange, além de ferramentas proprietárias como STATISTICA, SPSS Modeler, SAS Enterprise Miner, Microsoft Analysis Services, KXEN Infinite Insight e Oracle Data Mining. O estudo evidenciou a possibilidade de análise de dados educacionais com a aplicação de uma grande variedade de ferramentas que possibilitam resultados para suportar uma tomada de decisão mais eficiente no setor acadêmico.

O estudo de Rao, Swapna e Kumar (2018) mostrou que a Mineração de Dados permite que os usuários possam vislumbrar um cenário para tomada decisão a partir do conhecimento extraído dos bancos de dados. A mineração de dados é uma ferramenta utilizada de forma recorrente no campo da Educação nos últimos anos. Isto porque a mineração de dados educacionais revela-se um método eficaz para minerar o desempenho do aluno com base em vários parâmetros, possibilitando assim prever e analisar se um aluno será recrutado ou não a partir de sua capacitação na universidade. As previsões foram feitas usando os algoritmos de aprendizado de máquina J48, *Naive Bayes, Random Forest e Random Tree* nas ferramentas Weka e Multiple Linear Regression, regressão logística binomial, particionamento recursivo e árvore de regressão, árvore de inferência condicional e rede neural com algoritmos no estúdio R. O estudo indicou como resultado de cada abordagem que foram comparadas em relação aos seus níveis de desempenho e precisão por meio de análise gráfica.

A mineração de dados educacionais emergiu como uma “ferramenta altamente eficiente para identificar relações ocultas entre os dados acadêmicos, permitindo

assim antecipar o desempenho acadêmico dos estudantes” (YAĞCI, 2022, p. 11). Em função de múltiplos fatores influenciarem o desempenho do aluno de formas não lineares, este campo torna-se mais atraente aos pesquisadores. Também por conta da ampla disponibilidade de conjuntos de dados educacionais, o que contribui para o aumento do interesse dos pesquisadores, especialmente em contextos de aprendizado *online*.

No trabalho de Khan e Ghosh (2020), foi apresentada uma revisão sistemática de 140 estudos de EDM sobre o desempenho do aluno na aprendizagem em sala de aula. O foco dessa revisão se voltou à identificação dos preditores, métodos utilizados para tal identificação, tempo e objetivo da predição. Trata-se de uma pesquisa sistemática de estudos de EDM que considera a aprendizagem em sala de aula no transcurso do tempo. Os resultados indicaram que os pesquisadores alcançam uma eficiência de previsão significativa durante o período do curso enfocado na análise.

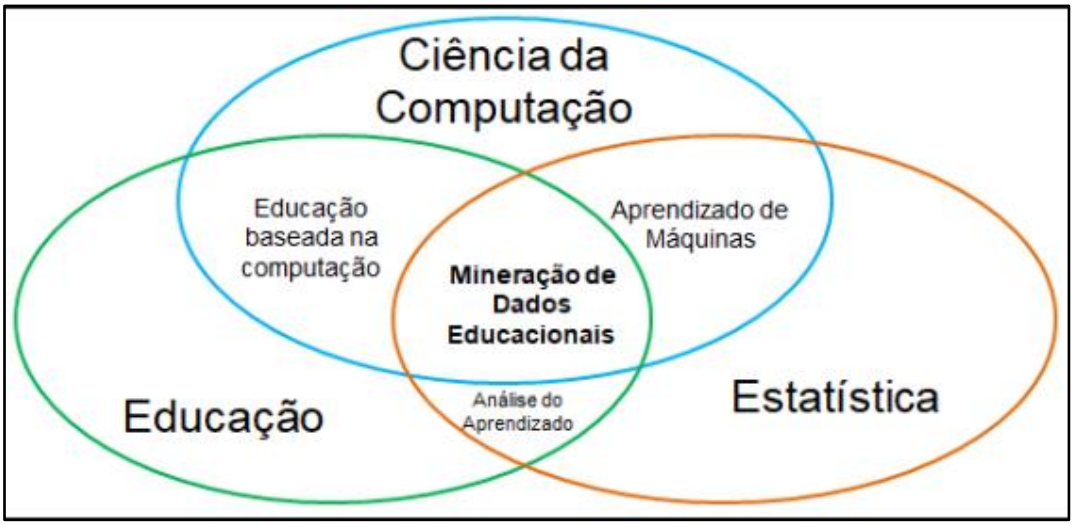
Diferentes pesquisas foram realizadas com foco na previsão do desempenho do aluno, a fim de apoiar seu desenvolvimento no curso. Muitas instituições estão focadas em melhorar o desempenho e a qualidade da educação; e isso pode ser alcançado utilizando técnicas de mineração de dados para analisar e prever o desempenho dos alunos e determinar possíveis fatores que podem afetar suas notas finais. No trabalho de Injadat, *et al.,* (2020) são analisados dois conjuntos de dados diferentes em dois estágios diferentes da evolução no curso (20% e 50%, respectivamente). A análise de recursos fornece informações sobre a natureza dos diferentes recursos considerados e ajuda na escolha dos algoritmos de aprendizado de máquina e seus parâmetros. Além disso, neste trabalho os autores propuseram uma abordagem sistemática baseada no índice de *Gini* e valor para a seleção de um aprendiz adequado a partir da combinação de seis algoritmos de aprendizado de máquina em potencial. Após os testes, os autores obtiveram resultados experimentais que demonstraram que os modelos propostos alcançam alta precisão e baixa taxa de falsos positivos nos estágios analisados em ambos os conjuntos de dados. Como resultado, foi possível delinear com maior clareza uma visão da evolução acadêmica do aluno e assim proporcionar decisões mais assertivas sobre as melhorias a serem implantadas.

Para Masethe *et al.* (2021), a introdução de métodos de Mineração de Dados Educacionais apresenta uma nova perspectiva para os tomadores de decisão de gestão educacional em instituições de ensino, que possibilita melhorar a precisão da previsão do desempenho acadêmico do aluno. No entanto, a diversidade e complexidade das diferentes técnicas de mineração de dados educacionais representam um enorme desafio, resultando em incertezas na tomada de decisões de gestão educacional. A contribuição dos autores para a área foi desenvolver uma abordagem de Sistema de Recomendação (RS) com a necessidade de um sistema de apoio à decisão para orientar o processo de escolha da técnica apropriada de Mineração de Dados Educacionais (EDM). O *framework* JCOLIBRI proposto pelos autores é utilizado na construção do Case-Based Reasoning - Recommender System (CBR-RS), que permite uma interface para um usuário não especialista definir uma consulta com base no domínio do problema.

Mahdi (2020) combinou o uso de sistema de gerenciamento de *e-learning online* (Moodle) com ferramentas de mineração de dados para melhorar o desempenho e a eficácia das maneiras de aprender e ensinar, aplicando para tanto os dados diários coletados das instituições educacionais. Seu trabalho mostra a relevância da mineração de dados educacionais, principalmente em relação às pesquisas de aplicações da mineração de dados e desenvolvimento de novas técnicas de mineração de dados no setor educacional.

A Mineração de Dados Educacionais (EDM), pode ser visualizada como a união das principais áreas da tecnologia voltada à educação, como: Ciência da Computação, Educação e Estatística, a união destas três áreas pode-se obter uma ampla análise da área de educação baseada em computação (*e-learning*), mineração de dados e aprendizado de máquina e análise do aprendizado (ROMERO; VENTURA, 2013; STOLL, 2019), conforme demonstrado na Figura 1.

Figura 1: Combinação das áreas da aplicação do EDM



Fonte: Romero e Ventura (2013)

A Figura 1 representa a interseção da mineração de dados educacionais com as demais áreas relacionadas. Essa representação visual evidencia a natureza multidisciplinar da mineração de dados educacionais e destaca as conexões entre diferentes campos de conhecimento. Ao centro da imagem está a mineração de dados educacionais, que é o foco principal. Ela envolve a aplicação de técnicas de mineração de dados e aprendizado de máquina para extrair conhecimento dos dados educacionais, buscando melhorar a tomada de decisões e aprimorar o ambiente educacional (ROMERO; VENTURA, 2013).

Ao redor da mineração de dados educacionais são apresentadas outras áreas de estudo intrinsecamente ligadas a ela. O aprendizado de máquina é uma dessas áreas, representando a utilização de algoritmos e modelos para o desenvolvimento de sistemas capazes de aprender e tomar decisões baseadas nos dados. A educação na computação refere-se ao campo que investiga a aplicação da ciência da computação e da tecnologia no contexto educacional. Nesse sentido, a mineração de dados educacionais desempenha um papel crucial na obtenção de informações valiosas para aprimorar os processos de ensino e aprendizagem.

Em resumo, as correlações expressadas por Romero e Ventura (2013) ilustram a integração da mineração de dados educacionais com diferentes áreas relacionadas, destacando a relevância dessa abordagem para o avanço da educação e o aprimoramento dos processos educacionais por meio da análise de dados.

2.4 Métodos e Técnicas de Mineração de Dados Educacionais

Este tópico inicia-se com a exposição de estudos que aplicaram com êxito diferentes métodos e técnicas de mineração de dados educacionais, para assim estabelecer o conjunto de métodos e técnicas considerado nesta pesquisa, quais sejam: K-means, Hierarchical clustering (Agrupamento Hierárquico), Gaussian Mixture Models (GMM), Agglomerative clustering (Agrupamento Aglomerativo). Que se referem a algoritmos de clusterização. Algoritmos de aprendizado de máquina não supervisionado usado para agrupar dados em grupos ou clusters com base em sua similaridade. O objetivo destes, é encontrar estruturas intrínsecas nos dados, agrupando-os de forma que observações semelhantes fiquem no mesmo cluster, enquanto observações diferentes são atribuídas a clusters diferentes. Na sequência são apresentadas estas ferramentas de mineração de dados educacionais consideradas para a execução dos experimentos previstos para esta pesquisa.

De acordo com Ahmed Fahim, (2021) K-means é um método de agrupamento altamente eficiente para lidar com conjuntos de dados de grande escala. No entanto, é mais adequado para conjuntos de dados que possuem agrupamentos com tamanhos semelhantes e formas convexas. Ele é usado para identificar a estrutura de clusters em um conjunto de dados, onde os objetos dentro do mesmo cluster são altamente semelhantes entre si, enquanto os objetos de diferentes clusters são altamente dissimilares. Porém, é necessário descobrir o valor de K para ter a visão exata de quantos clusters deve ser aplicado no algorítimo K-means. Em seu estudo o autor aplicou uma mescla de dois algoritmos de agrupamento; DBSCAN e k-means, para analisar se há um melhor desempenho do algorítimo com o uso conjunto e obteve resultados significativos.

Segundo Sinaga e Yang (2020) o algoritmo K-means é considerado o método de agrupamento mais conhecido e usado na literatura e possui diversas extensões, em seu estudo as autoras propuseram um novo algoritmo, não supervisionado, de clusterização k-means o *“Uk-means”*, que possui a finalidade de encontrar o número ideal de clusters sem fornecer nenhuma inicialização e seleção de parâmetro, sendo assim de forma automática. Resultados experimentais e comparações realmente demonstram esses bons aspectos do algoritmo de agrupamento de médias.

Para Govender e Sivakumar (2020) a Clusterização é uma técnica de análise exploratória de dados usada desde 1930, porém, obteve maior popularidade na década de 60, ela busca investigar estruturas subjacentes nos dados e possui uma utilização em grande escala em estudos com dados atmosféricos e poluição do ar. Dentro desta técnica, existem algoritmos que fazem este agrupamento de dados, de forma a agrupar por similaridade/semelhança, um destes algoritmos é o Agglomerative clustering (Agrupamento Aglomerativo), apesar de bem explorada na prática, há uma carência de revisões que abordam esta temática.

BRIGGS, *et al*., (2020) em seu trabalho apresentou uma modificação para FL (aprendizado federado) introduzindo uma etapa de clusterização hierárquica (FL+HC) para separar clusters de clientes pela similaridade de suas atualizações locais com o modelo de junta global. Uma vez separados, os clusters são treinados de forma independente e em paralelo em modelos especializados. Seus resultados mostraram como é possível aumentar a precisão do conjunto analisados, com uso do algoritmo Hierarchical clustering (Agrupamento Hierárquico) juntamente a suas extensões.

2.4.1 Métodos de Mineração de Dados Educacionais - K-means

K-means: É um algoritmo amplamente utilizado que busca particionar os dados em k clusters, onde k é um número predefinido. Ele atribui cada observação ao cluster mais próximo, com base nas médias dos atributos.

2.4.2 Métodos de Mineração de Dados Educacionais - Hierarchical Clustering

Hierarchical Clustering (Agrupamento Hierárquico): É um método que cria uma estrutura de árvore hierárquica de clusters, onde os clusters podem ser unidos ou divididos sucessivamente com base em sua similaridade. Ele pode ser aglomerativo (começando com cada ponto como um cluster separado e mesclando-os) ou divisivo (começando com um único cluster contendo todos os pontos e dividindo-o).

2.4.3 Métodos de Mineração de Dados Educacionais - Gaussian Mixture Models (GMM)

Gaussian Mixture Models (GMM): É um modelo probabilístico que assume que os dados são gerados por uma mistura de distribuições Gaussianas. Ele tenta estimar

os parâmetros dessas distribuições e atribuir as observações a cada componente do modelo.

2.4.4 Métodos de Mineração de Dados Educacionais - Agglomerative Clustering

Agglomerative Clustering (Agrupamento Aglomerativo): É um método de agrupamento hierárquico que começa com cada ponto como um cluster separado e, em seguida, mescla sucessivamente os clusters mais próximos uns dos outros até que todos os pontos estejam em um único cluster.

Esses são apenas alguns exemplos de algoritmos de clusterização. Cada algoritmo tem suas próprias vantagens, desvantagens e aplicações adequadas. A escolha do algoritmo certo depende das características dos dados, do objetivo da análise e das suposições subjacentes que podem ser feitas sobre os dados. Há uma grande diversidade de algoritmos de aprendizado de máquina e ferramentas de pesquisa aplicáveis na Mineração de Dados (MD). Com o crescimento da Mineração de Dados Educacionais (EDM) verificado nos últimos anos, novas ferramentas foram geradas para suprir esta demanda de aplicações (MAHAJAN; SAINI, 2020). A seguir são apresentadas soluções para a mineração de dados educacionais aplicáveis a diferentes problemas e situações da gestão acadêmica em instituições de ensino.

2.4.5 Ferramentas de Mineração de Dados Educacionais

O RapidMiner é uma plataforma de software de ciência de dados desenvolvida pela empresa de mesmo nome que fornece um ambiente integrado para a preparação de dados, aprendizado de máquina, aprendizado profundo, mineração de texto e análise preditiva. Esta plataforma pode ser acessada no endereço https://rapidminer.com/platform/educational/.

O KNIME é uma plataforma livre e de código aberto de análise de dados, construção de relatórios e integração de dados. O KNIME integra vários componentes para aprendizado de máquina e mineração de dados por meio de seu conceito de pipelining modular. Esta plataforma pode ser acessada no endereço https://www.knime.com/.

O Orange é um kit de ferramentas de visualização de dados de código aberto, aprendizado de máquina e mineração de dados. Ele apresenta um *front-end* de programação visual para a análise exploratória e visualização interativa de dados qualitativos. Este kit pode ser acessado no endereço https://orangedatamining.com/.

SPSS é um aplicativo de cunho científico. Originalmente o nome era acrônimo de *Statistical Package for the Social Sciences* (pacote estatístico para as ciências sociais). Ele ajuda a eliminar a lacuna entre a ciência de dados e a compreensão de dados. Este aplicativo pode ser acessado no endereço https://www.ibm.com/br-pt/spss.

O Apache Spark é um mecanismo multilíngue para executar engenharia de dados, ciência de dados e aprendizado de máquina em clusters ou máquinas de nó único. Este mecanismo pode ser acessado no endereço https://spark.apache.org/.

2.5 Gestão Acadêmica

Para Santos (2018), a gestão acadêmica está relacionada ao entendimento amplo do funcionamento da universidade e também ao conhecimento adquirido a partir do modelo de trabalho do corpo docente ao analisar processos e decisões institucionais; ao mesmo tempo que são consideradas atividades de gestão que possuam algum envolvimento de alunos junto ao corpo docente.

Segundo Gusso *et al.* (2020), um fator dificultador da gestão acadêmica nos últimos anos foi a chegada da pandemia de Covid-19. Isto porque foi necessária uma mudança brusca e rápida na tomada de decisão por parte dos gestores de universidades, a fim de garantir a continuidade do processo de ensino-aprendizagem.

De acordo com Ramos (2019), no âmbito das IES o planejamento na gestão acadêmica é ferramenta de suma importância para um bom desenvolvimento e entregas no processo de ensino-aprendizagem. Segundo Kayser, Silva e Braga (2016), um bom planejamento na gestão acadêmica da IES pode assegurar melhores políticas educacionais e nortear estratégias para a instituição de ensino, a fim de assim alcançar suas metas e consolidar sua missão institucional.

São vários os momentos nos quais os gestores e corpo docente da instituição de ensino se juntam para compor um conjunto de ações voltadas à condução dos

cursos na instituição. Destaca-se, em especial, o momento que antecede o ano letivo, uma vez que neste período é traçado um novo planejamento que tem por base o histórico de períodos anteriores, visando assim proporcionar melhorias para o próximo período letivo. Pesquisa conduzida por Ruiz (2019) criou um inovador modelo *Canvas* voltado à identificação das necessidades da instituição de ensino que propicia o planejamento acadêmico, reajustando o plano de ensino e melhorando a integração do corpo docente. Seus resultados demonstraram que o ambiente fornecido pelo modelo *Canvas* criado ajudou a suportar a gestão acadêmica ao possibilitar uma visão mais ampla dos processos de ensino-aprendizagem.

2.6 Pesquisa em Bases de Dados de Publicações Científicas sobre a Temática Abordada nesta Pesquisa

No Quadro 1 é apresentada a estratégia de busca aplicada nas bases de dados de publicações científicas selecionadas para compor este estudo. As buscas foram realizadas nas bases Web of Science, IEEE e Scopus. Também são apresentadas as palavras-chave consideradas e a *string* de busca (Inteligência Artificial - *Artificial intelligence,* Mineração de Dados - *Data mining,* Mineração de Dados Educacionais - *Educational data mining,* Gestão Acadêmica - *Academic management,* Ensino Superior - *University education*) de busca aplicada nas bases. A pesquisa foi realizada em abril de 2023 e o intervalo de tempo estipulado foram os últimos cinco anos (2019 a 2023).

Quadro 1: Estratégia de busca nas bases de dados de publicações científicas

BASE	Publicações dos 5 últimos anos	Apenas Artigos	Artigos relacionados (Área Educação)
WEB OF SCIENCE	52.132	1.253	176
IEEE XPLORE	77.485	32.092	304
SCOPUS	6.455	2.424	328

Fonte: Autora (2023)

O Quadro 1 exibe os resultados de busca nas duas bases de dados selecionadas (Web of Science, IEEE e Scopus), a fim de trazer da literatura os trabalhos mais relevantes e próximos à temática abordada nesta dissertação. As palavras-chave consideradas foram: Artificial intelligence; Data mining; Educational data mining; Academic management e University education, com critério de exclusão por área de conhecimento, para maior aproximação ao tema considerado.

No Quadro 2, são apresentados 10 trabalhos com maior aderência e alinhamento ao tema proposto por esta pesquisa coletados na busca na base de dados Web of Science.

Quadro 2: Trabalhos identificados com maior aderência à temática abordada na pesquisa

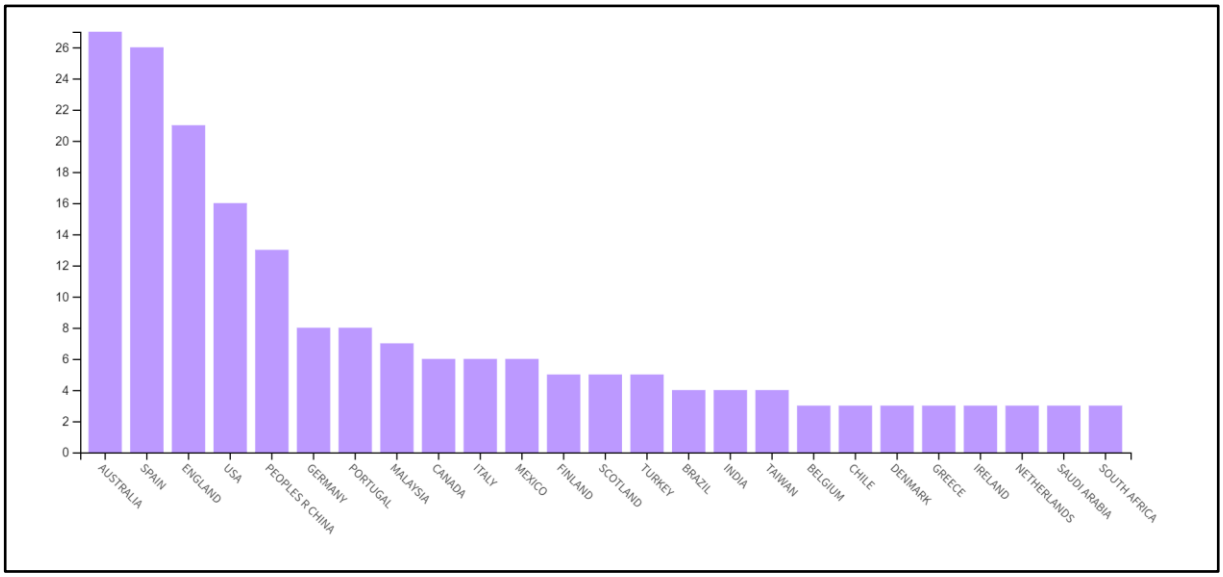
Título da Publicação	Autores, ano da publicação	Resumo
Affordances and challenges of artificial intelligence in K-12 education: a systematic review	Crompton, Helen; Jones, Mildred, V; Burke, Diane (2022)	Este artigo nos traz a evolução nas interações da IA no ensino superior, desde experiências dos alunos até a gestão de cursos por meio de ferramentas de diagnóstico e conteúdo da disciplina.
Predicting academic success in higher education: literature review and best practices	Alyahyan, Eyman; Dustegor, Dilek (2020)	Este estudo visa fornecer um conjunto passo a passo de diretrizes para educadores dispostos a aplicar técnicas de mineração de dados educacionais para prever o sucesso do aluno.
Artificial intelligence applications in Latin American higher education: a systematic review	Salas-Pilco, Sdenka Zobeida; Yang, Yuqin (2022)	Neste trabalho é demonstrado que os aplicativos de IA ajudam a abordar questões educacionais importantes (por exemplo, detectar alunos em risco de abandono) e, assim, contribuir para uma melhor gestão e qualidade de cursos.
Artificial intelligence in online higher education: A systematic review of empirical research from 2011 to 2020	Ouyang, Fan; Zheng, Luyi; Jiao, Pengcheng (2022)	Neste trabalho incluem previsão do status de aprendizado, desempenho ou satisfação, recomendação de recursos, avaliação automática e melhoria da experiência de aprendizado;
Student engagement in online learning in Latin American higher education during the COVID-19 pandemic: A systematic review	SALAS-PILCO, Sdenka Zobeida; YANG, Yuqin; ZHANG, Zhe (2022)	Este trabalho se trata de uma revisão sistemática da literatura sobre o envolvimento do aluno na aprendizagem online no ensino superior. Onde sintetiza as descobertas sobre o engajamento estudantil em instituições de ensino superior latino-americanas durante a pandemia do COVID-19.
Dropout at university. Variables involved on it	Lorenzo-Quiles, O; Galdon-Lopez, S; Lendinez-Turon, A (2023)	Este trabalho tem por objetivo: analisar a satisfação do aluno, especificar as causas da evasão e determinar os autores mais adequados sobre a evasão por meio da literatura e de diferentes bancos de dados. Após análise concluiu-se que cinco componentes principais estariam por trás da evasão

		universitária: adaptação, personalidade, nível socioeconômico, relacionamento professor-aluno e qualidade na educação universitária.
Causes of academic dropout in higher education in Andalusia and proposals for its prevention at university: A systematic review	de la Cruz-campos, JC; Victoria-Maldonado, JJ; Martinez-Domingo, JA; Campos-Soto, MN (2023)	Este estudo traz a realidade do desinteresse acadêmico por partes de alunos ingressos ao ensino superior, concluiu-se que as causas mais frequentes do abandono universitário estão associadas ao baixo desempenho acadêmico, fraco apoio social no novo ambiente acadêmico, baixo nível socioeconômico, pessimismo e falta de motivação, juntamente com outros fatores menos significativos, como a má relacionamento com professores, falta de vocação, incompatibilidade de trabalho e desempenho acadêmico anterior.
Programmes targeting student retention/success and satisfaction/experience in higher education: A systematic review	Eather, N; Mavilidi, MF; Sharp, H; Parkes, R (2022)	O objetivo desta revisão sistemática foi relatar o sucesso de intervenções ou programas realizados em universidades especificamente visando melhorar os resultados dos alunos com base em dados quantitativos publicados.
A Review of Using Machine Learning Approaches for Precision Education	Luan, H; Tsai, CC (2021)	Neste artigo, revisamos sistematicamente 40 estudos empíricos sobre educação de precisão baseada em aprendizado de máquina. Os resultados mostraram que a maioria dos estudos se concentrou na previsão de desempenho ou evasão de aprendizagem e foram realizados em ambientes de aprendizagem online ou combinados.
Evaluating the role of bursaries in widening participation in higher education: a review of the literature and evidence	Kaye, N (2021)	Este trabalho teve o objetivo de examinar a eficácia das bolsas para capacitação de jovens a fim de compensar a desvantagem financeira, e também do impacto que sua oferta nas atitudes dos jovens e na integração dos estudantes na universidade. As descobertas destacaram a persistência de disparidades socioeconômicas na participação no ensino superior e, mesmo entre os alunos de baixa renda que frequentavam, enfatizavam os desafios contínuos que enfrentaram para se adequar à vida universitária.

Fonte: Autora (2023)

A Figura 2 apresenta o ranking das publicações obtidas na base Web of Science, revelando a produção científica em diferentes países. A seguir os dados dos cinco primeiros colocados nesse ranking.

Figura 2: Trabalhos publicados na Web of Science sobre a temática abordada

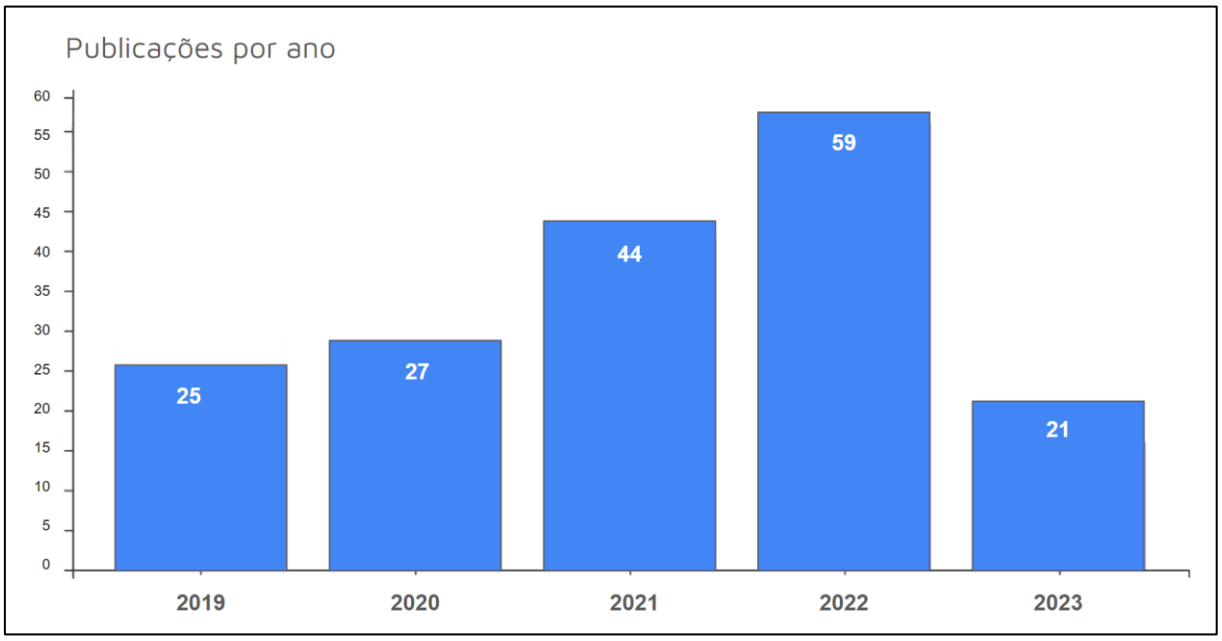


Fonte: Web of Science (2023)

A Austrália se destaca com 27 artigos publicados (15,34% do total de publicações). Logo em seguida, a Espanha aparece com 26 artigos (14,77%), seguida pela Inglaterra em terceiro lugar, com 21 artigos (11,93%). Na quarta posição verifica-se os Estados Unidos da América, com 16 artigos publicados (9,09%). Por fim, a China aparece em quinto lugar, com 13 artigos (7,39%). O Brasil vem surgir apenas em décimo quinto lugar, com 4 artigos (2,27% das publicações).

Na Figura 3 é apresentada a quantidade de publicações voltadas à temática Inteligência Artificial, Mineração de Dados e Mineração de Dados Educacionais na Gestão Acadêmica do Ensino Superior, no período de 2019 a abril de 2023.

Figura 3: Trabalhos analisados por ano de publicação na base Web of Science (2019 a abril/2023)

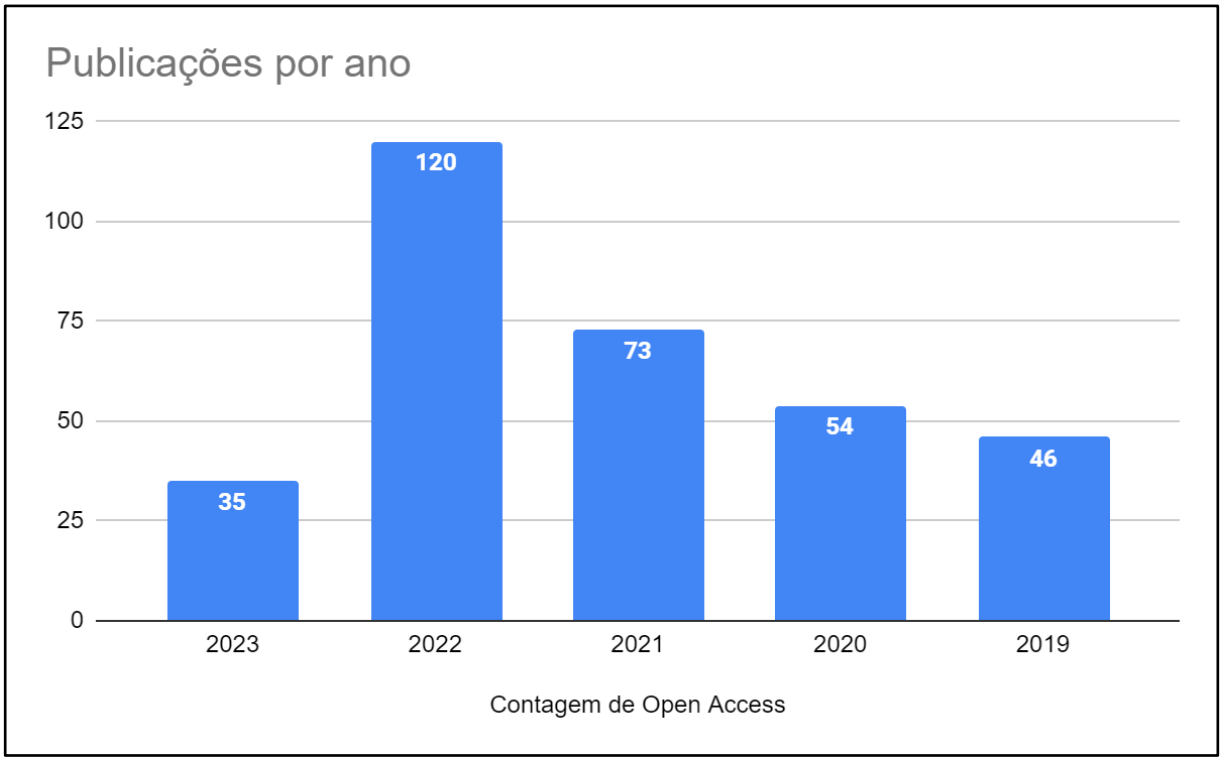


Fonte: Web of Science (2023)

Conforme indicado na figura anterior, nota-se significativo avanço na quantidade publicações nos últimos cinco anos, considerando-se que o período de corte é abril/2023, ou seja, o ano de 2023 está restrito ao primeiro quadrimestre. As áreas de pesquisa encontradas nos trabalhos identificados têm despertado interesse crescente de acadêmicos, profissionais e pesquisadores devido à sua relevância e potencial impacto na gestão e no aprimoramento dos processos educacionais. Isto porque as aplicações identificadas nas pesquisas encontras permitem a análise de grandes volumes de dados e a identificação de padrões relevantes, que podem ser utilizados para melhorar a tomada de decisões e o planejamento estratégico nas instituições de ensino superior.

Na Figura 4 é apresentada uma análise feita sob a luz de outra base de dados (Scopus) de trabalhos científicos na temática enfocada nesta dissertação, considerando-se o mesmo período de 2019 a abril/2023.

Figura 4: Trabalhos analisados por ano de publicação (Scopus - 2019 a 2023)



Fonte: Scopus (2023)

Como é possível verificar na Figura 4, há tendência de crescimento nas publicações relacionadas à temática abordada nesta dissertação também na base Scopus, considerando-se que foram considerados trabalhos publicados até o primeiro quadrimestre de 2023. Tal tendência demonstra o interesse crescente na utilização dessas tecnologias para propiciar melhorias na qualidade da educação, aumentar a eficiência dos processos acadêmicos e contribuir para a tomada de decisões informadas nas instituições de ensino superior.

3. MÉTODO E MATERIAIS DE PESQUISA

Neste capítulo são apresentados o método e materiais de pesquisa considerados nesta dissertação. Para tanto, o foco maior é destinado à apresentação da estrutura e características da solução automatizada delineada que pode ser viabilizada por meio de um protótipo para a execução desta pesquisa. O conceito de protótipo é apresentado, bem como sua aplicabilidade em pesquisas e trabalhos científicos desenvolvidos em diferentes áreas de estudo. Por fim, é descrita a aplicação da solução automatizada a partir de um protótipo desenvolvido voltado à gestão de dados acadêmicos a fim de promover melhorias na gestão acadêmica de grandes massas de dados gerados em Instituições de Ensino Superior.

A metodologia desta dissertação é caracterizada como exploratória, quantitativa, aplicada e experimental, tendo sido realizada por meio de experimentação e análises dos resultados obtidos. A pesquisa exploratória busca investigar e compreender fenômenos pouco explorados ou ainda não totalmente compreendidos na área de mineração de dados educacionais. É uma abordagem inicial que visa fornecer percepções e gerar prognósticos para pesquisas futuras (HUNTER; MCCALLUM, 2019).

A pesquisa quantitativa envolve a coleta e análise de dados numéricos e estatísticos relacionados às realizações acadêmicas dos alunos. Ela se concentra na aplicação de métodos estatísticos para entender as relações e padrões entre as variáveis estudadas, conforme indicado por Bryman (2004) e Mehrad e Zangeneh (2019). A pesquisa aplicada tem o objetivo de utilizar os resultados obtidos para resolver problemas práticos ou tomar decisões relevantes de diversos campos (GOLDSMITH, 2021). No contexto educacional ao aplicar a mineração de dados, o foco é melhorar o desempenho acadêmico dos alunos e otimizar os processos educacionais.

Por fim, a pesquisa experimental envolve a manipulação de variáveis e a aplicação de intervenções controladas para verificar o impacto na previsão de realizações acadêmicas. Nesse tipo de pesquisa, é possível realizar testes e comparações com grupos de controle, permitindo uma abordagem sistemática para avaliar os resultados obtidos (MEHRAD; ZANGENEH, 2019).

A priori foi realizada pesquisa bibliográfica seguindo as seguintes etapas metodológicas para identificar, selecionar e analisar fontes de informação relevantes relacionadas ao tema de interesse.

1) Definição do tema: o primeiro passo foi definir o tema e pergunta de pesquisa, para delimitar o escopo da pesquisa e a identificar as palavras-chave relevantes.
2) Busca de fontes: com base no tema definido, foi realizado buscas em bases de dados acadêmicas (Web of Science, IEEE e Scopus) e repositórios de periódicos. as palavras-chave identificadas foram usadas para encontrar publicações relacionadas ao assunto.
3) Seleção de fontes: após a busca, foi avaliado as fontes encontradas, analisando o título, resumo e palavras-chave para verificar a relevância.
4) Leitura e análise: análise do material feito anotações sobre os principais pontos, conceitos e conclusões relevantes.
5) Discussão e conclusões: com base nas informações encontradas na pesquisa bibliográfica, foram discutidos os resultados, contextualizando as descobertas em relação ao tema de pesquisa e apresentando suas conclusões com base no conjunto de fontes consultadas.

Após esta etapa, desenvolveu-se um modelo conceitual detalhado da solução proposta, delineando os passos necessários para sua construção. Posteriormente, foram empregadas técnicas de agrupamento *(clustering)* de aprendizado de máquina não supervisionado para avaliar a viabilidade da abordagem proposta. Ao finalizar essa etapa, os resultados obtidos foram analisados, culminando na criação de um guia de implementação que oferece orientações práticas para a adoção da solução, juntamente com recomendações sobre a aplicação da Mineração de Dados Educacionais (EDM). Esse guia e as recomendações visam aprimorar a tomada de decisão nos cursos superiores, proporcionando um valioso suporte aos gestores acadêmicos.

3.1 Protótipo

Um protótipo é uma versão preliminar ou modelo inicial de um produto ou sistema criado com o objetivo de testar e validar sua funcionalidade, *design* e usabilidade antes de serem desenvolvidas versões definitivas (PINTO; MOTA, 2020). De acordo com Lee (2018), esta é uma fase de converter ideias em uma forma física, que seja rápida e barata para que possa vivenciar e interagir com a ideia, e no processo, aprendendo e desenvolvendo. Um protótipo é uma “representação concreta e tangível de uma ideia, que permite aos designers e desenvolvedores testar, experimentar e explorar diferentes soluções e possibilidades”. O autor destaca ainda que o protótipo não precisa ser perfeito ou completo, mas sim suficiente para permitir que os testes e validações sejam realizados de forma eficiente e eficaz (LAZUARDI; SUKOCO, 2019, p. 4.).

Para Pressman (2021), o protótipo é uma técnica importante para reduzir os riscos e incertezas no processo de desenvolvimento de uma solução de software, permitindo que os usuários e *stakeholders* envolvidos possam interagir com o sistema de forma antecipada e assim fornecer *feedback* valioso para o desenvolvimento subsequente. Ou seja, no contexto desta pesquisa a solução automatizada ora vislumbrada é uma ferramenta essencial ao processo de desenvolvimento de produtos e sistemas, permitindo que os *designers* e desenvolvedores testem e validem diferentes soluções antes de investirem recursos significativos no desenvolvimento final.

3.2 Aplicação da Solução

Nesta pesquisa foi desenvolvido um modelo de solução projetado com a ferramenta Astah Community, que é uma aplicação voltada a realizar modelos de diagramas UML, tais como: caso de uso, diagrama de classes e diagrama de componentes. Nesta seção são apresentadas as funcionalidades do diagrama de caso de uso para cada usuário envolvido na solução delineada, o diagrama de classes e suas atribuições e, por fim, o diagrama de componentes para a criação da solução automatizada de mineração de dados educacionais para suporte à gestão acadêmica de cursos superiores.

3.2.1 Diagrama de Caso de Uso

Na Linguagem de Modelagem Unificada (UML), o diagrama de caso de uso resume os detalhes dos usuários do seu sistema (também conhecidos como atores) e as interações deles com o sistema. O diagrama de caso de uso UML é ideal para: representar as metas de interações entre sistemas e usuários, definir e organizar requisitos funcionais no sistema, especificar o contexto e os requisitos do sistema e modelar o fluxo básico de eventos no caso de uso (PEIXOTO, 2016).

Nesta pesquisa o diagrama de caso de uso é utilizado para demonstrar a interação entre os atores e a solução automatizada proposta com suas funcionalidades. Assim, o diagrama de caso de uso desta pesquisa apresenta a projeção, definição da ferramenta e mapeamento do comportamento da solução a ser desenvolvida. Neste caso, a solução possui três atores com acessos distintos à solução a ser desenvolvida, conforme indicado no quadro 3.

Quadro 3: Atores e descrição de acesso

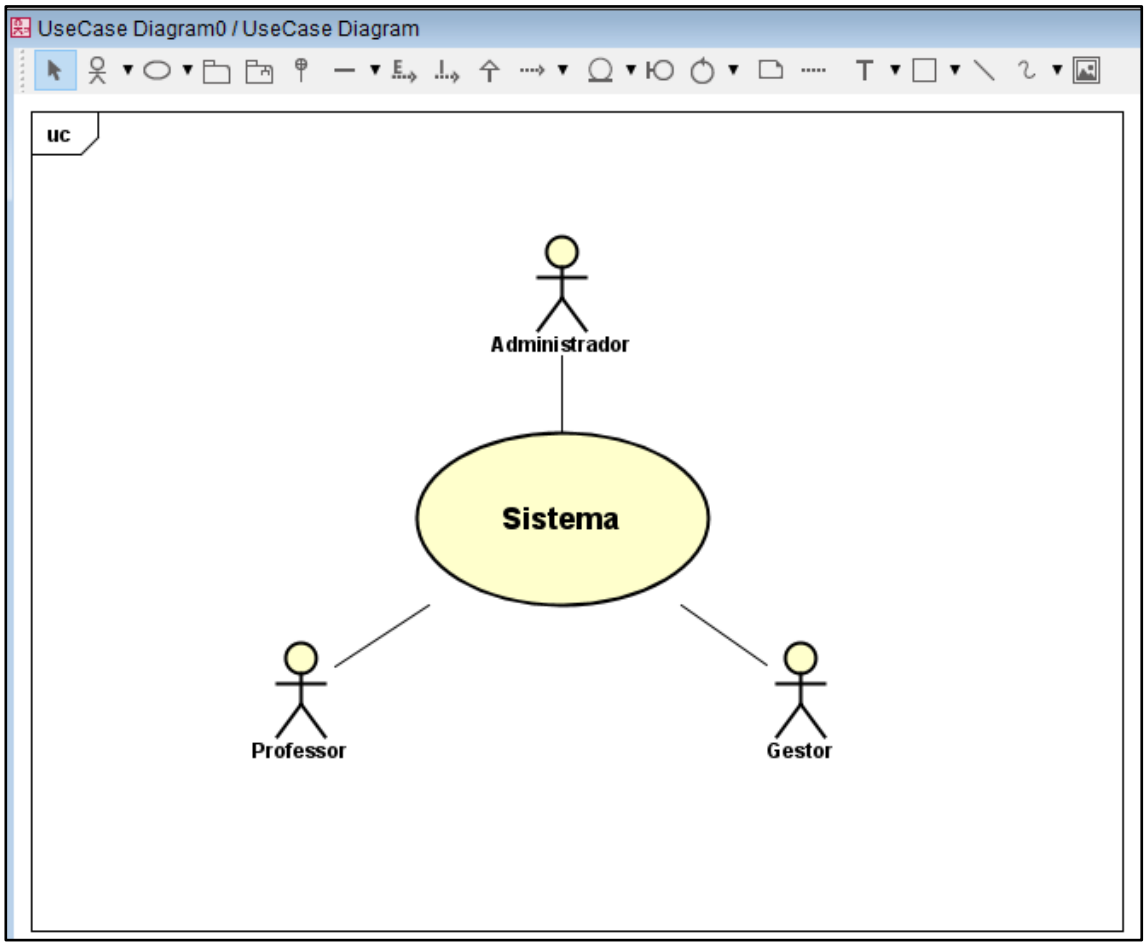
Ator	**Descrição de Acesso**
Professor	Usuário que fornece dados de desempenho dos alunos, tais como nota, créditos cumpridos, disciplinas, etc.
Gestor	Usuário, com acesso aos dados e informações disponíveis, descritas por meios de gráficos que apresentam o desempenho individual do aluno, por curso ou geral para acompanhamento e gestão dos resultados e proposição de melhorias.
Administrador	Usuário responsável pela administração do sistema, geralmente a cargo do departamento de TI da instituição de ensino.

Fonte: Autora (2023)

No Quadro 3 foram apresentadas as atribuições do perfil de cada ator responsável com acesso ao sistema, sendo o professor o responsável pela entrada dos dados de evolução do aluno, tais como notas, faltas e comentários pertinentes à evolução do aluno durante seu tempo de curso. Já o Gestor terá a função de gerar relatórios e gráficos de forma regular para a análise do perfil do aluno e suas individualidades durante o tempo em que este se mantiver no curso para ao final de cada semestre obter um panorama da situação de cada aluno. Por fim, o Administrador será responsável pelo funcionamento do sistema, oferecendo suporte

necessário aos usuários e promovendo atualizações quando necessárias. Estes atores são os principais agentes de funcionamento para que a solução a ser elaborada possa conter dados suficientes e assim cumprir com os requisitos planejados. O diagrama de caso de uso com a indicação dos atores envolvidos é exposto na Figura 5, que foi elaborada no software Astah.

Figura 5: Diagrama de Caso de Uso - Atores

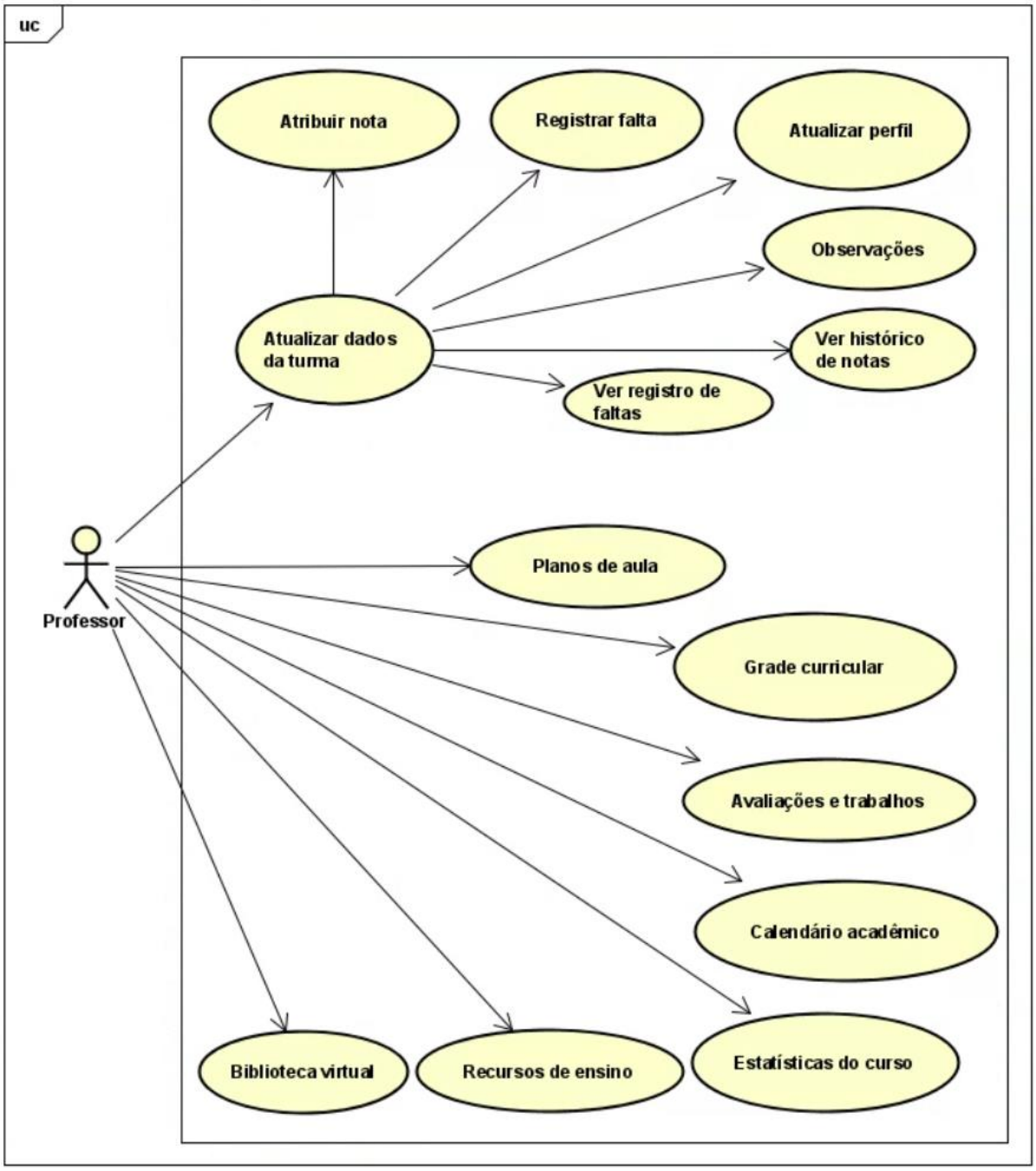


Fonte: Autora (2023)

Os atores ilustrados no diagrama de caso de uso são: professor, gestor e administrador, sendo que cada qual possui funções distintas, porém todos se relacionam com o sistema. A seguir na Figura 6 é possível visualizar os atores com seus respectivos relacionamentos e as principais funcionalidades individuais, a partir de maior detalhamento do diagrama de caso de uso já exposto.

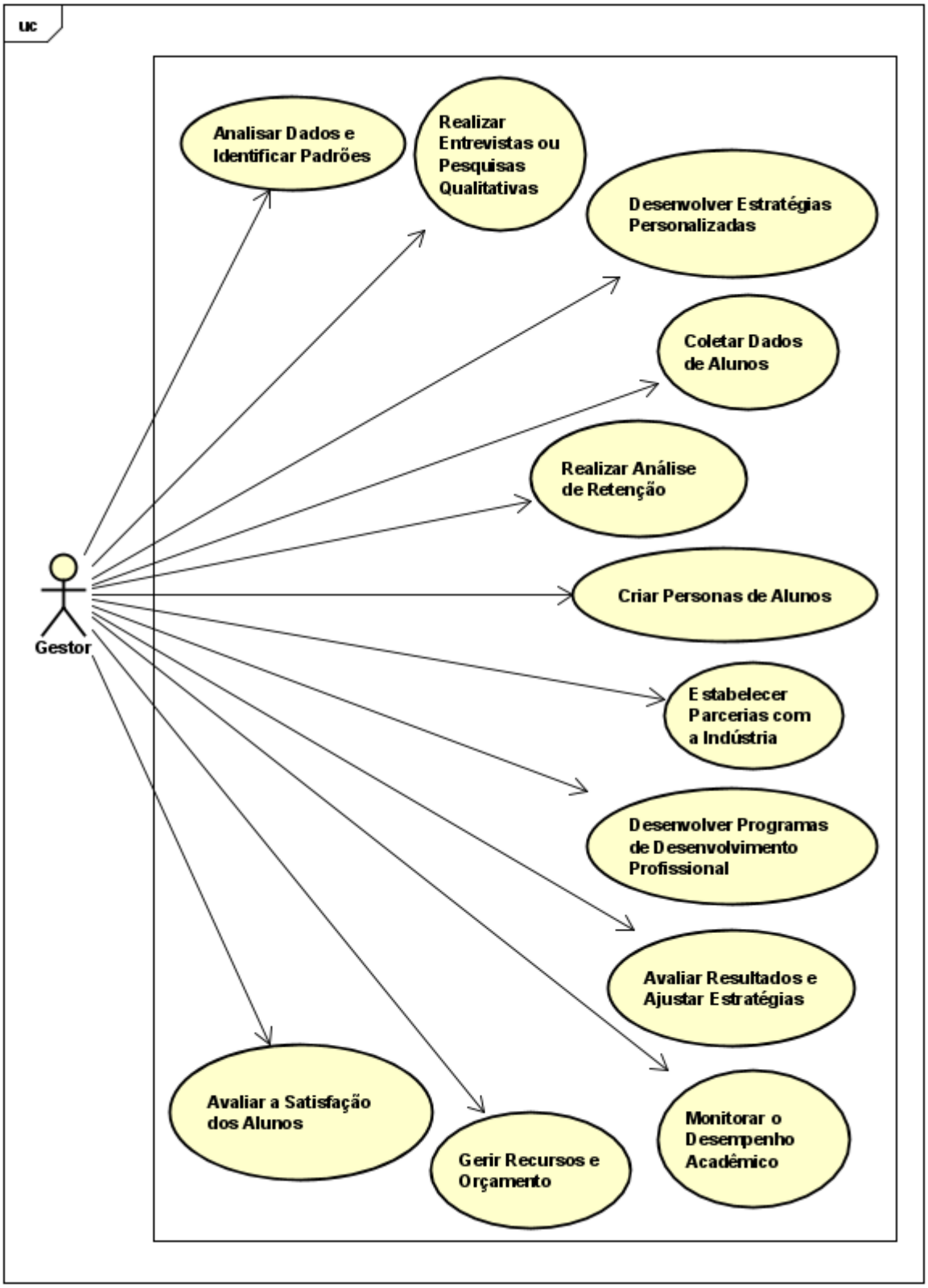
Figura 6: Diagrama de Caso de Uso – Relacionamentos e funcionalidades dos atores

a) Professor



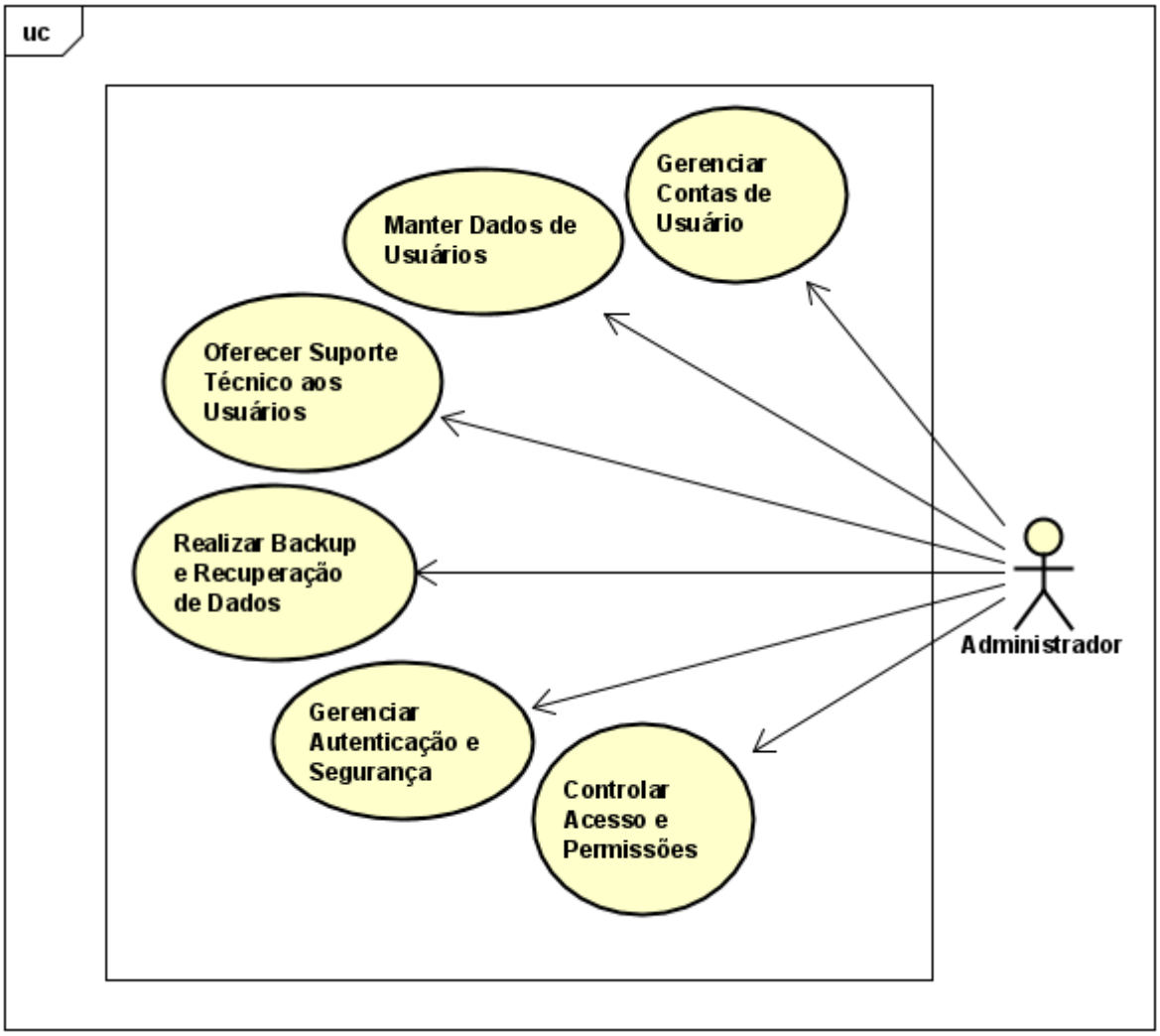
Fonte: Autora (2023)

b) Gestor



Fonte: Autora (2023)

c) Administrador



Fonte: Autora (2023)

Na figura 6 os atores ilustrados são apresentados com sua principal função, a saber: a) Professor - responsável pela alimentação dos dados e conceitos relacionados aos alunos; b) Gestor – responsável por prover análises baseadas nos dados adquiridos do desenvolvimento dos alunos durante o semestre e c) Administrador - responsável por manter o sistema funcionando e passível a novos cadastros de gestores e docentes. A seguir são indicadas em detalhe as definições de cada atribuição de casos de uso dos três diferentes agentes:

a) Professor

● "Atualizar dados da turma": este caso de uso representa a ação do professor de atualizar os dados de cada aluno desta turma. Ele pode ter atributos como "aluno", "disciplina", "nota", “presença” e descrever o fluxo de eventos envolvendo:

● “Atribuir nota”: este caso de uso permite que o professor atribua uma nota a uma atividade, trabalho, prova ou qualquer outro componente de avaliação do aluno. Ao clicar nesse botão, o professor pode inserir a nota correspondente e salvá-la no sistema.

● “Registrar falta”: este caso de uso permite que o professor registre a falta de um aluno em uma aula, seminário, laboratório ou qualquer outra atividade acadêmica. Ao clicar nesse botão, o professor pode selecionar a data e a atividade específica em que o aluno faltou e salvar a informação.

● “Atualizar perfil”: este caso de uso permite que o professor atualize o perfil do aluno com informações relevantes, como observações, feedbacks, sugestões ou qualquer outra informação pertinente. Ao clicar nesse botão, o professor pode inserir as informações necessárias e salvar as alterações no perfil do aluno.

● “Observações”: este caso de uso permite que o professor faça uma anotação referente a turma, como uma atividade, ou tema que será trabalhado com a turma posteriormente, ou até um feedback adicional para salvar no sistema.

● “Ver histórico de notas”: este caso de uso permite que o professor acesse o histórico de notas do aluno, visualizando todas as notas atribuídas ao longo do período letivo. Isso pode ajudar o professor a ter uma visão geral do desempenho do aluno em diferentes atividades.

● “Ver registro de faltas”: este caso de uso permite que o professor acesse o registro de faltas do aluno, visualizando todas as faltas registradas ao longo do período letivo. Isso pode auxiliar o professor a identificar padrões de ausência e tomar as medidas necessárias.

"Planos de aula": este caso de uso representa a ação do professor acessar os planos de aula do curso, fornecendo um cronograma detalhado das aulas, tópicos abordados, materiais de leitura e atividades relacionadas.

"Grade curricular": este caso de uso representa a ação do professor acessar a grade curricular do curso, visualizando as disciplinas que compõem o currículo, suas descrições, carga horária e requisitos;

"Avaliações e trabalhos": este caso de uso permite exibir as avaliações e trabalhos atribuídos aos alunos do curso, permitindo ao professor visualizar as datas de entrega, critérios de avaliação e registrar as notas correspondentes.

"Calendário acadêmico": este caso de uso permite ao professor acessar o calendário acadêmico do curso, visualizando datas importantes, como períodos de matrícula, prazos de entrega de trabalhos, datas de provas e feriados.

"Estatísticas do curso": este caso de uso permite ao professor gerar relatórios estatísticos relacionadas ao desempenho do curso, como média geral das notas, taxa de aprovação, distribuição de notas, entre outros indicadores relevantes.

"Recursos de ensino": este caso de uso permite ao professor ter acesso a recursos de ensino específicos do curso, como apresentações de slides, materiais didáticos, vídeos, listas de leitura recomendada, entre outros.

"Biblioteca virtual": este caso de uso permite ao professor acessar à biblioteca virtual, onde poderá pesquisar e encontrar materiais acadêmicos relevantes para o curso, como artigos científicos, livros e periódicos.

b) Gestor

"Coletar Dados de Alunos": este caso de uso permite ao gestor coletar dados detalhados dos alunos, incluindo informações pessoais, histórico acadêmico, preferências de cursos, interesses, habilidades, entre outros. Isso pode envolver a criação de formulários de pesquisa ou a integração com sistemas de registro acadêmico para coletar os dados necessários.

"Realizar Entrevistas ou Pesquisas Qualitativas": este caso de uso permite ao gestor realizar entrevistas ou pesquisas qualitativas com os alunos para obter

percepções mais aprofundados sobre seus perfis. Isso pode envolver a condução de entrevistas individuais, grupos focais ou pesquisas de satisfação para entender melhor suas necessidades, expectativas e motivações.

"Analisar Dados e Identificar Padrões": este caso de uso permite ao gestor analisar os dados coletados dos alunos e identificar padrões e tendências significativas. Isso pode envolver o uso de técnicas de análise de dados, mineração de dados e aprendizado de máquina para descobrir informações valiosas sobre os perfis de alunos, como segmentos distintos, características comuns, preferências de cursos, entre outros.

"Criar Personas de Alunos": este caso de uso permite ao gestor criar personas de alunos, que são perfis fictícios que representam grupos de alunos com características semelhantes. Isso envolve a consolidação dos dados coletados e a criação de personas que encapsulam as principais características, necessidades, objetivos e motivações dos diferentes grupos de alunos.

"Desenvolver Estratégias Personalizadas": este caso de uso permite ao gestor desenvolver estratégias personalizadas para atender às necessidades específicas de cada grupo de alunos. Isso pode envolver o design de programas acadêmicos, ações de suporte ao aluno, iniciativas de engajamento, estratégias de retenção e outras abordagens adaptadas aos perfis identificados.

"Avaliar Resultados e Ajustar Estratégias": este caso de uso permite ao gestor avaliar os resultados das estratégias implementadas e ajustar as abordagens com base no feedback dos alunos e nos indicadores de desempenho. Isso envolve o acompanhamento dos resultados das estratégias implementadas, a análise dos dados de desempenho (Estatísticas do curso) e a realização de ajustes para melhor atender às necessidades dos alunos.

"Monitorar o Desempenho Acadêmico": este caso de uso permite ao gestor monitorar o desempenho acadêmico dos alunos em cada curso. Isso pode envolver a análise de notas, médias, taxas de sucesso, evasão e outros indicadores para identificar tendências e tomar medidas adequadas para melhorar o desempenho dos alunos.

"Realizar Análise de Retenção": este caso de uso permite ao gestor realizar análises de retenção de alunos em cada curso. Isso envolve a identificação de fatores

que influenciam a retenção, como taxas de abandono, razões para a desistência e implementação de estratégias para aumentar a retenção de alunos.

"Gerir Recursos e Orçamento": este caso de uso permite ao gestor gerir os recursos e o orçamento relacionados a cada curso. Isso inclui a alocação de recursos, planejamento financeiro, análise de custos, identificação de necessidades de investimento e tomada de decisões sobre o uso eficiente dos recursos disponíveis.

"Estabelecer Parcerias com a Indústria": este caso de uso permite ao gestor estabelecer parcerias com a indústria ou instituições externas relevantes para cada curso. Isso pode envolver a identificação de oportunidades de colaboração, estabelecimento de acordos de estágio ou parcerias de pesquisa, e promoção de programas conjuntos para enriquecer a experiência dos alunos e melhorar a empregabilidade.

"Desenvolver Programas de Desenvolvimento Profissional": este caso de uso permite ao gestor desenvolver programas de desenvolvimento profissional para os alunos em cada curso. Isso envolve a identificação de necessidades de treinamento, criação de workshops, palestras ou seminários relevantes, e facilitação do desenvolvimento de habilidades práticas e competências necessárias para o mercado de trabalho.

"Avaliar a Satisfação dos Alunos": este caso de uso permite ao gestor avaliar a satisfação dos alunos em cada curso. Isso pode envolver a realização de pesquisas de satisfação, coleta de feedback dos alunos e análise dos resultados para identificar áreas de melhoria e implementar ações para melhorar a experiência dos alunos.

c) Administrador

"Gerenciar Contas de Usuário": este caso de uso permite ao administrador criar, modificar e remover contas de usuário no sistema. O administrador pode criar novos usuários, atribuir permissões de acesso, configurar senhas e gerenciar as informações de perfil de cada usuário.

"Controlar Acesso e Permissões": este caso de uso permite ao administrador controlar o acesso e as permissões dos usuários no sistema. Isso envolve atribuir diferentes níveis de acesso e permissões com base nas funções e responsabilidades

de cada usuário. O administrador pode garantir que apenas usuários autorizados tenham acesso a determinados recursos e funcionalidades do sistema.

"Manter Dados de Usuários": este caso de uso permite ao administrador manter e atualizar os dados dos usuários no sistema. Isso pode incluir informações pessoais, como nome, endereço de e-mail, número de telefone, bem como outras informações relevantes, como departamento, cargo e área de atuação.

"Gerenciar Autenticação e Segurança": este caso de uso permite ao administrador gerenciar a autenticação e a segurança do sistema. O administrador pode configurar políticas de senha, implementar autenticação de dois fatores, monitorar atividades suspeitas e garantir a conformidade com as práticas de segurança da informação.

"Oferecer Suporte Técnico aos Usuários": este caso de uso permite ao administrador oferecer suporte técnico aos usuários do sistema. O administrador pode ajudar os usuários com problemas de acesso, fornecer orientações sobre a utilização do sistema, solucionar problemas técnicos e responder a dúvidas relacionadas às contas de usuário.

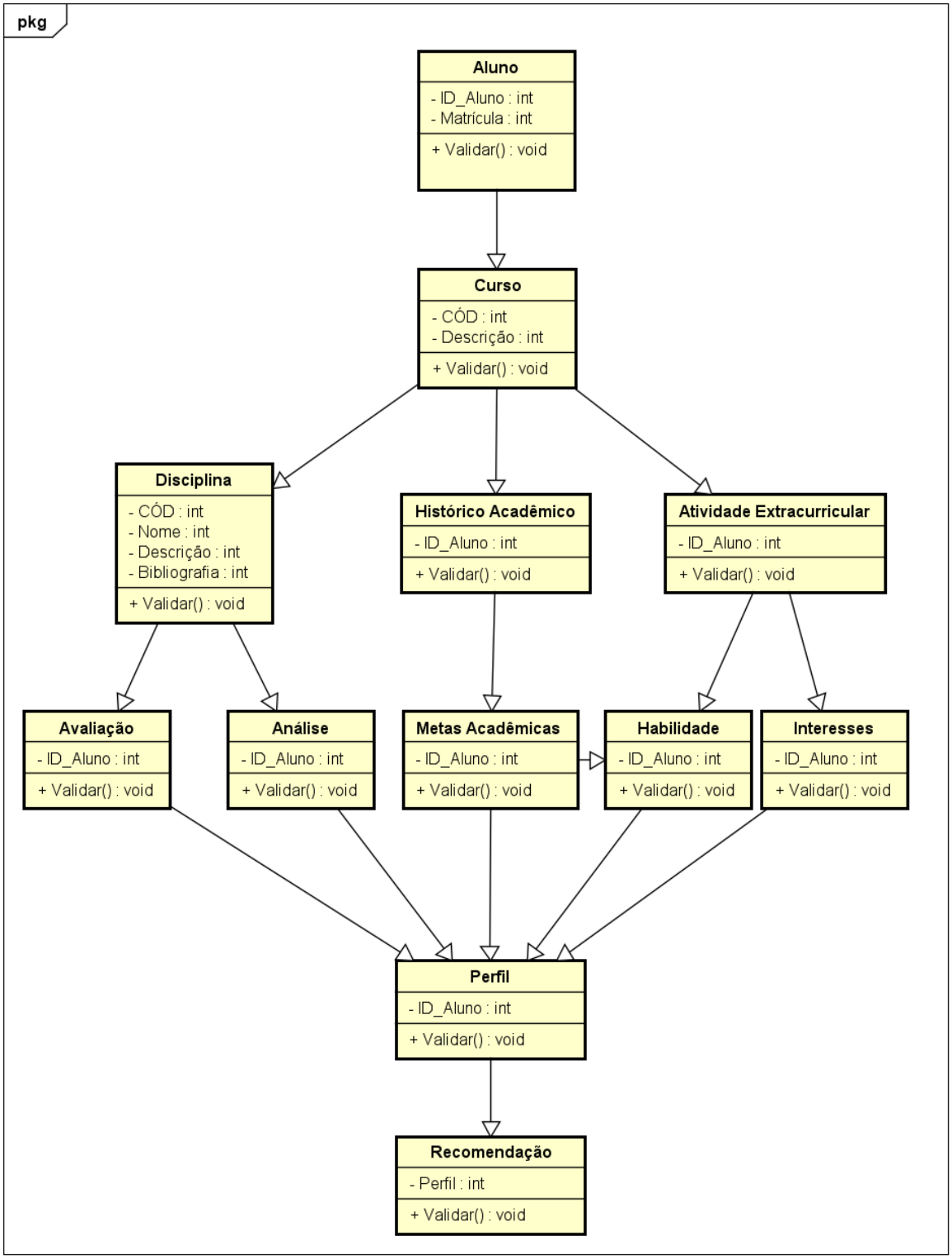
"Realizar Backup e Recuperação de Dados": este caso de uso permite ao administrador realizar backups regulares dos dados do sistema e garantir a recuperação eficiente em caso de falhas ou perdas de dados. O administrador pode implementar políticas de backup, monitorar a integridade dos dados e executar procedimentos de recuperação quando necessário.

3.2.2 Diagrama de Classe

O próximo diagrama utilizado para estruturar a solução desenvolvida nesta pesquisa é o Diagrama de Classes. Este diagrama contém as classes que caracterizam os objetos da solução a ser desenvolvida. As classes são obtidas a partir da análise do Diagrama de Casos de Uso e representam os componentes de interação com o sistema e seus relacionamentos. Uma classe é representada por um retângulo sólido com três partes: a primeira para o nome da classe, outra para os atributos da classe (que podem ser vistos como características da classe) e a terceira para a declaração das operações definidas para a classe (BACURAU, *et al.*, 2022). Na Figura

7 é exposto o diagrama de classes com a notação UML desenvolvido no software Astah para exposição das classes da solução a ser desenvolvida.

Figura 7: Diagrama de Classe



Fonte: Autora (2023)

O Diagrama de Classes expõe as características dos atores mediante suas funções anteriormente descritas no tópico anterior. Este diagrama possui um formato simples com a notação UML para ilustrar as principais funcionalidades da solução automatizada a ser desenvolvida. A seguir as classes e suas respectivas explicações:

Classe "Aluno": representa o aluno e contém informações relacionadas ao perfil do aluno, como nome, matrícula, idade, gênero, histórico acadêmico, interesses, habilidades, preferências de cursos, entre outros atributos importantes para a análise do perfil.

Classe "Disciplina": representa as disciplinas individuais oferecidas nos cursos. Ela pode conter informações sobre o nome da disciplina, carga horária, conteúdo programático, pré-requisitos, entre outros atributos relevantes para a análise do perfil do aluno.

Classe "Curso": representa os cursos oferecidos pela instituição acadêmica. Ela pode conter informações sobre o nome do curso, carga horária, disciplinas, requisitos, objetivos e outros atributos relevantes para o perfil do aluno.

Classe "Perfil": representa o perfil do aluno e contém informações sobre as características e preferências do aluno. Ela pode incluir atributos como habilidades, interesses, estilo de aprendizagem, metas acadêmicas, preferências de metodologia de ensino, entre outros aspectos que definem o perfil individual do aluno.

Classe "Avaliação": representa as avaliações realizadas ao longo do curso. Ela pode conter informações sobre as notas do aluno em cada avaliação, critérios de avaliação, ponderação das notas e outros atributos relevantes para a análise do desempenho acadêmico do aluno.

Classe "Análise": representa a análise do perfil do aluno. Ela pode conter métodos e atributos relacionados à análise dos dados do aluno, identificação de padrões, geração de percepções, cálculos estatísticos e outros processos que contribuem para o diagnóstico do perfil do aluno.

Classe "Recomendação": representa as recomendações personalizadas com base no perfil do aluno. Ela pode conter informações sobre cursos recomendados,

disciplinas complementares, atividades extracurriculares, recursos de apoio ao aluno e outras sugestões alinhadas ao perfil do aluno.

Classe "Habilidade": representa as habilidades específicas relacionadas a cada curso. Ela pode conter informações sobre as habilidades esperadas para os alunos em cada curso, níveis de proficiência, histórico de desenvolvimento de habilidades e outros atributos primordiais para a análise do perfil do aluno.

Classe "Atividade Extracurricular": representa as atividades extracurriculares disponíveis para os alunos. Ela pode conter informações sobre clubes, grupos de estudo, projetos, eventos, estágios e outras oportunidades extracurriculares que podem influenciar o perfil do aluno.

Classe "Metas Acadêmicas": representa as metas acadêmicas estabelecidas pelos alunos. Ela pode conter informações sobre as metas individuais de cada aluno, como conquistar uma bolsa de estudos, alcançar um determinado desempenho acadêmico, participar de programas de intercâmbio, entre outras metas importantes para a análise do perfil do aluno.

Classe "Histórico Acadêmico": representa o histórico acadêmico do aluno. Ela pode conter informações sobre o desempenho do aluno em termos de notas, frequência, progressão nos cursos, histórico de conclusão de disciplinas e outros registros acadêmicos essenciais para a análise do perfil do aluno.

Classe "Interesses": representa os interesses individuais do aluno. Ela pode conter informações sobre os interesses acadêmicos, áreas de estudo preferidas, tópicos de pesquisa, projetos pessoais e outros aspectos que refletem os interesses do aluno e influenciam seu perfil.

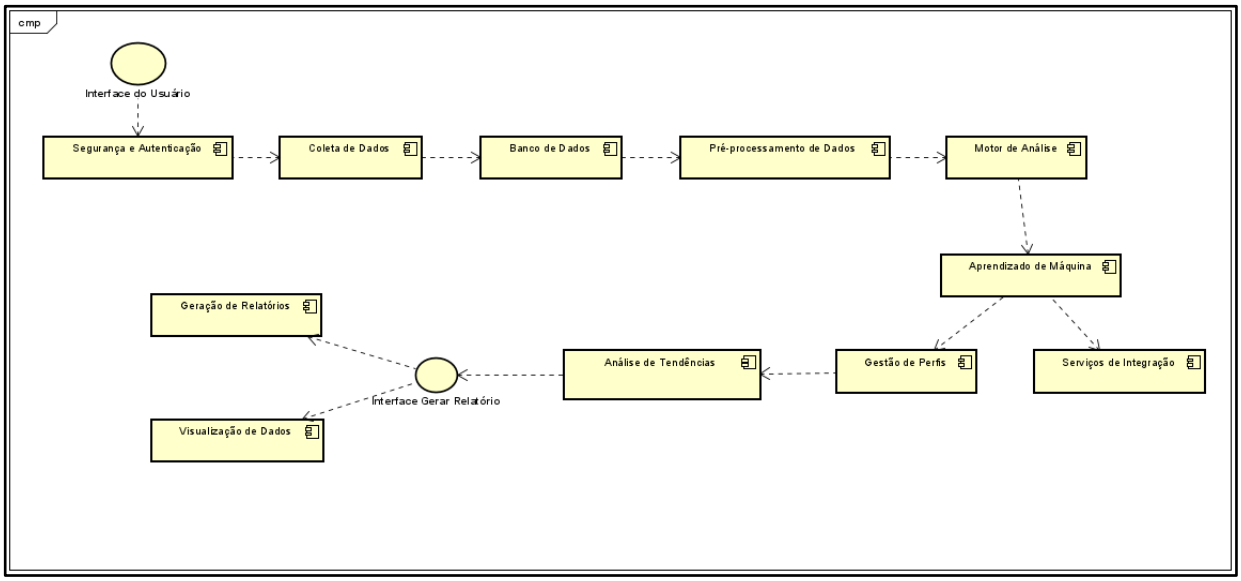
3.2.3 Diagrama de Componentes

O próximo diagrama a compor a estruturação da solução a ser desenvolvida é o Diagrama de Componentes. Este diagrama contém um tipo de estrutura que faz parte da linguagem de modelagem unificada (UML). Ele é usado em conjunto com outros diagramas da UML, como diagramas de classes, diagramas de sequência e

diagramas de atividades. O diagrama de componentes fornece uma visão abrangente do sistema e auxilia no desenvolvimento de software, criando de forma visual a relação entre os componentes, portas e interfaces. É possível utilizar diagramas de componentes para modelar sistemas de software em um alto nível ou para mostrar componentes em um nível de pacote mais baixo (BETTIN; GERALDI; OLIVEIRA JR., 2018).

Um componente é uma unidade modular e independente de software que encapsula funcionalidades relacionadas, podendo ser uma biblioteca, um módulo, um serviço, uma classe ou qualquer outra unidade de software reutilizável. Os componentes são representados no diagrama por retângulos com o nome do componente (SHCHERBAN, *et al.*, 2021). Na Figura 8 é exposto o diagrama de componentes da solução a ser desenvolvida.

Figura 8: Diagrama de Componentes



Fonte: Autora (2023)

O Diagrama de Componentes apresentado é baseado nas funções dos Atores e Classes anteriormente descritas, que visam prever e/ou diagnosticar o perfil de aluno, oferecendo um maior entendimento ao Gestor, promovendo assim melhoria da tomada de decisão. Os componentes e respectivas representações são explicados a seguir.

Componente "Interface do Usuário": representa a interface do sistema que permite aos usuários interagir com as funcionalidades relacionadas ao diagnóstico do perfil do aluno. Ele pode incluir elementos como telas, formulários, botões e outros elementos de interface necessários para a entrada e exibição de dados relacionados ao perfil do aluno.

Componente "Banco de Dados": representa o banco de dados que armazena os dados relacionados aos alunos, cursos, disciplinas, avaliações e outros elementos relevantes para o diagnóstico do perfil do aluno. Ele pode incluir tabelas, relacionamentos entre as entidades e consultas para recuperar e atualizar os dados necessários para a análise do perfil.

Componente "Coleta de Dados": representa os mecanismos e processos para coletar dados relevantes sobre os alunos. Pode incluir integração com sistemas de gestão acadêmica, formulários de feedback, questionários, dados de desempenho acadêmico, informações socioeconômicas e outras fontes de dados utilizadas para construir o perfil do aluno.

Componente "Pré-processamento de Dados": trata do pré-processamento dos dados coletados antes da análise. Pode incluir etapas como limpeza de dados, normalização, tratamento de valores ausentes, transformação de variáveis e outras técnicas para garantir a qualidade e a adequação dos dados para a análise do perfil do aluno.

Componente "Motor de Análise": representa o motor de análise responsável por processar os dados do aluno e realizar a análise para obter o diagnóstico do perfil. Ele pode conter algoritmos, modelos de aprendizado de máquina, cálculos estatísticos e outras técnicas de análise de dados que são aplicadas para identificar padrões e gerar percepções sobre o perfil do aluno.

Componente "Serviços de Integração": representa os serviços de integração com outros sistemas ou fontes de dados externas que podem ser utilizados para enriquecer a análise do perfil do aluno. Por exemplo, pode incluir integrações com sistemas de aprendizado de máquina pré-treinados, bancos de dados de referência ou serviços de terceiros que fornecem informações adicionais sobre os alunos.

Componente "Geração de Relatórios": representa a funcionalidade responsável por gerar relatórios e visualizações dos resultados da análise do perfil do aluno. Ele pode incluir recursos para criar gráficos, tabelas, dashboards e outros elementos visuais que comunicam com entendimento obtido sobre o perfil do aluno.

Componente "Segurança e Autenticação": representa os mecanismos de segurança e autenticação necessários para proteger os dados relacionados ao perfil do aluno. Ele pode incluir recursos como criptografia, controle de acesso, gerenciamento de permissões e outras medidas de segurança para garantir a confidencialidade e integridade dos dados do aluno.

Componente "Aprendizado de Máquina": engloba os algoritmos e modelos de aprendizado de máquina utilizados na análise do perfil do aluno. Pode incluir técnicas de agrupamento (clusterização), classificação, regressão, processamento de linguagem natural e outras abordagens de aprendizado de máquina para identificar padrões e extrair compreensões do perfil do aluno.

Componente "Visualização de Dados": concentra-se na representação visual dos dados e resultados da análise do perfil do aluno. Pode incluir gráficos interativos, dashboards, mapas de calor, infográficos e outras formas de visualização que permitem aos gestores e tomadores de decisão compreender e explorar os esclarecimentos gerados.

Componente "Gestão de Perfis": lida com a gestão dos perfis dos alunos, incluindo a criação, atualização e exclusão de perfis. Pode incluir funcionalidades para armazenar e recuperar perfis, rastrear mudanças ao longo do tempo e gerenciar a privacidade e segurança dos dados do aluno.

Componente "Análise de Tendências": concentra-se na identificação e análise de tendências de desempenho acadêmico e comportamento dos alunos ao longo do tempo. Pode incluir análises estatísticas, modelagem preditiva e técnicas de séries temporais para identificar padrões e prever comportamentos futuros dos alunos.

3.3 Descrição da Solução

Neste tópico é descrito o modelo conceitual para a execução da solução automatizada proposta nesta dissertação, que consiste em uma ferramenta para análises de dados acadêmicos visando formular um prognóstico de perfil de aluno ingressante em cursos superiores em apoio à gestão acadêmica. A ferramenta delineada na solução automatizada em pauta permitirá ao gestor visualizar, por meio de análise de dados coletados no banco de dados acadêmicos de sua instituição, notas/conceitos, trabalhos desenvolvidos, projetos dos alunos e índices de aprovação, reprovação e trancamento de cada curso/disciplina. A partir destes dados, a solução de mineração de dados que subsidia a solução proposta é capaz de gerar perfis de potenciais candidatos/alunos mais alinhados a cada curso de graduação e pós-graduação, o que poderá contribuir para que a instituição ofereça cursos e disciplinas com maior eficiência.

A metodologia para cumprir os objetivos desta pesquisa foi a análise da literatura e desenvolvimento da solução automatizada com guia de orientações para o gestor. Primeiramente foi executada busca nos repositórios de trabalhos científicos para a sustentação teórica da temática abordada nesta pesquisa. Foram consultadas bases de dados científicas e bibliotecas digitais para identificar estudos e artigos relevantes sobre inteligência artificial, mineração de dados, mineração de dados educacionais e gestão acadêmica. Nos trabalhos identificados foram analisados os principais conceitos, teorias, técnicas e abordagens discutidos na literatura, a fim de embasar o desenvolvimento da solução automatizada, conforme indicação de De Sousa *et al.,* (2021).

A análise dos trabalhos encontrados nas bases científicas subsidiou o desenvolvimento da solução automatizada de mineração de dados educacionais, seguindo uma abordagem iterativa e incremental para sua evolução. Foi utilizado ambiente de programação (Python), juntamente com bibliotecas e ferramentas especializadas em mineração de dados e análise estatística, além do software *Tableau* para análise de dados. A solução automatizada foi projetada para coletar, processar e analisar os dados educacionais relevantes, tais como notas, frequência e informações socioeconômicas dos alunos, dentre outros. Algoritmos de aprendizado de máquina foram aplicados para identificação de padrões e tendências nos dados

em análise, permitindo assim a formulação de um prognóstico de perfil do aluno ingressante de interesse da instituição de ensino em cursos superiores.

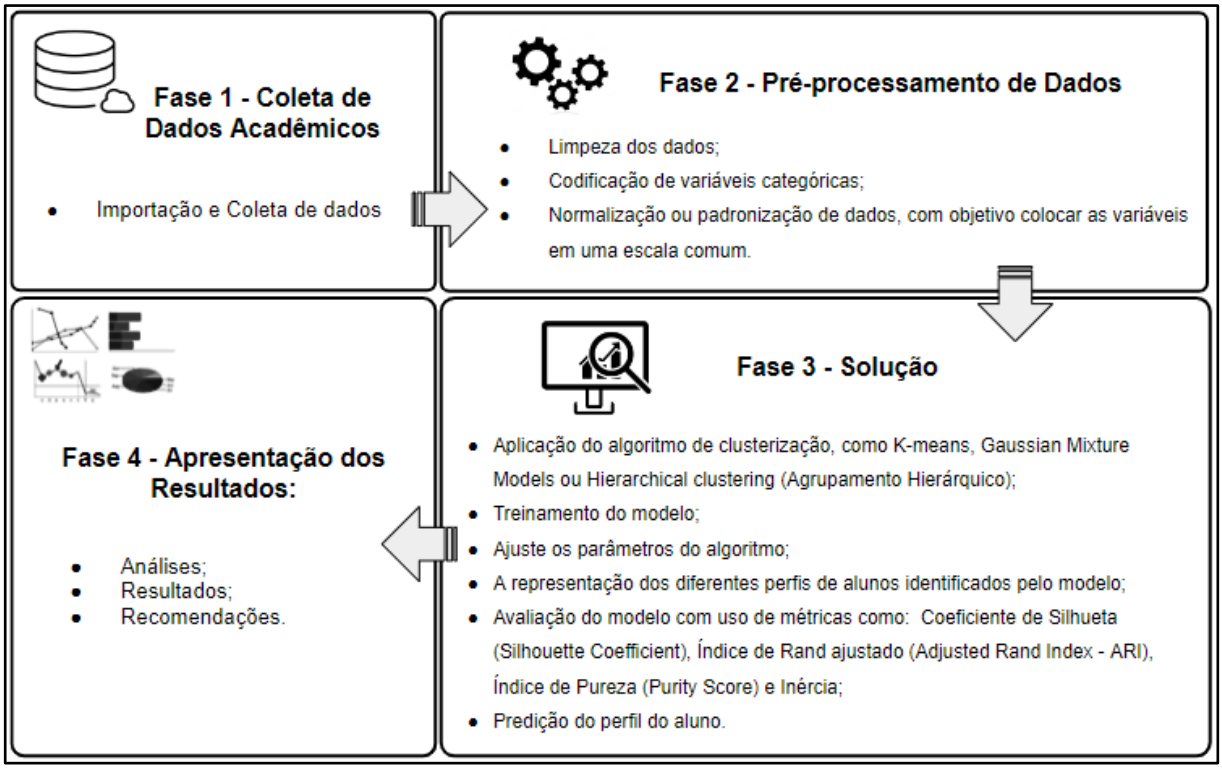
Por fim, foi realizada a validação da solução automatizada por meio de experimentos e avaliações, conforme indicação de pesquisa experimental estabelecida por Suave, *et al.,* (2021). Para tanto, foram selecionadas amostras de dados reais de alunos ingressantes em cursos superiores de pós-graduação e de dados reais de alunos e seus respectivos resultados acadêmicos ao final do curso. Desta forma, a solução automatizada foi desenvolvida com base nos dados desses dois bancos de dados, sendo os resultados comparados com informações já conhecidas sobre o desempenho e o perfil dos alunos.

As métricas de desempenho aplicadas nos experimentos foram Coeficiente de Silhueta (*Silhouette Coefficient*), Índice de Rand ajustado (*Adjusted Rand Index - ARI*), Índice de Pureza (*Purity Score*) e Inércia. Além dessas métricas, também é possível avaliar visualmente a separação dos *clusters* (agrumapentos) por meio de gráficos e análise de dados, como visualização de dispersão (*scatter plot*) ou visualização dos clusters em um espaço reduzido por meio de técnicas de redução de dimensionalidade, como PCA ou t-SNE.

Com base nos resultados obtidos e nas lições aprendidas durante o desenvolvimento e validação da solução, foram elaboradas telas para indicar o funcionamento da solução, do ponto de vista do gestor usuário da solução delineada. Também foi construído um guia de implementação dos requisitos para a implementação da solução automatizada de mineração de dados educacionais para suporte à gestão acadêmica de cursos superiores. Ele servirá como um recurso para gestores acadêmicos interessados em aplicar a mineração de dados educacionais para aprimorar a compreensão do perfil do aluno e auxiliar na tomada de decisões relacionadas à gestão acadêmica.

A Figura 9 apresenta o modelo conceitual do sistema idealizado nesta pesquisa experimental, fornecendo uma representação visual das etapas e componentes envolvidos na solução proposta. Esse modelo conceitual serve como referência para orientar o desenvolvimento e implementação da solução, garantindo a sua coerência e eficácia para a gestão acadêmica.

Figura 9: Modelo conceitual da solução



Fonte: Autora (2023)

Conforme indicado na Figura 9, as quatro fases da solução delineada são detalhadas a seguir:

Fase 1 - Coleta de Dados Acadêmicos: os dados adicionados ao longo do semestre, tais como notas, créditos de trabalhos, cursos prévios anteriores ao curso atual são selecionados e inseridos no banco de dados para análise.

Fase 2 - Pré-processamento de dados: antes de serem analisados, os dados selecionados e inseridos no banco de dados passam por um processo de tratamento. Essa etapa é essencial para ajustar os dados conforme as necessidades específicas da solução, resultando em uma forma otimizada para que os algoritmos possam realizar as operações voltadas à identificação e prognóstico de perfis de alunos. Durante o processo de tratamento e transformação dos dados, diferentes técnicas podem ser aplicadas, notadamente voltadas à limpeza dos dados, preenchimento de valores faltantes e normalização e padronização dos dados, dentre outras. Essas etapas visam garantir a integridade e consistência dos dados, além de prepará-los de

maneira adequada para a aplicação dos algoritmos de mineração de dados.

A limpeza dos dados envolve a remoção de registros duplicados, *outliers* e dados inconsistentes, eliminando assim dados incompletos, corrompidos ou irrelevantes que possam distorcer a análise a ser realizada. Já o preenchimento de valores faltantes é importante para evitar perdas de dados significativos e possibilitar uma análise mais completa, considerando-se um volume maior de dados para análise. A normalização e padronização dos dados pretendem colocar as variáveis em uma escala comum, garantindo que não haja viés ou desproporção nos resultados. Isto permite uma comparação mais precisa entre diferentes variáveis, além de facilitar a identificação de padrões e tendências. Ao final do processo de transformação e tratamento, os dados estarão prontos para serem utilizados pelos algoritmos de mineração de dados, conforme detalhado na próxima fase.

Fase 3 - Solução: em seguida, a solução automatizada foi projetada para coletar, processar e analisar os dados acadêmicos de interesse aos gestores das instituições de ensino. Isso pode envolver a criação de *scripts* ou programas em Python para extração dos dados de fontes diversas, como bancos de dados acadêmicos, planilhas eletrônicas ou sistemas de gerenciamento de aprendizagem. Os dados coletados podem incluir informações como notas, frequência, histórico acadêmico, informações socioeconômicas e qualquer outra variável que seja relevante para a análise do perfil do aluno ingressante em cursos superiores.

Em seguida foram aplicados algoritmos de aprendizado de máquina aos dados pré-processados. Esses algoritmos incluem técnicas de agrupamento (*clustering*) de aprendizado de máquina não supervisionado, como algoritmos K-means, *Hierarchical clustering* (Agrupamento Hierárquico), *Gaussian Mixture Models* (GMM) e *Agglomerative clustering (*Agrupamento Aglomerativo), dependendo dos objetivos específicos do prognóstico do perfil do aluno. Tais algoritmos são aplicados para analisar os padrões e tendências nos dados em processamento, buscando identificar características comuns entre os alunos ingressantes e auxiliando na formulação do prognóstico de cada perfil de aluno.

Ressalta-se ser importante conduzir testes rigorosos usando conjuntos de dados reais, comparando os resultados da solução automatizada com informações já conhecidas sobre os perfis dos alunos. Para tanto, métricas de desempenho, como

Coeficiente de Silhueta (*Silhouette Coefficient*), Índice de Rand ajustado (*Adjusted Rand Index* - ARI), Índice de Pureza (*Purity Score*) e Inércia foram aplicadas para avaliar a qualidade do prognóstico e a eficácia da solução automatizada. A abordagem iterativa e incremental, combinada com o uso de um ambiente de programação adequado, bibliotecas especializadas e algoritmos de aprendizado de máquina permitirá o desenvolvimento de uma solução automatizada de mineração de dados educacionais capaz de formular prognósticos do perfil do aluno ingressante em cursos superiores de forma confiável.

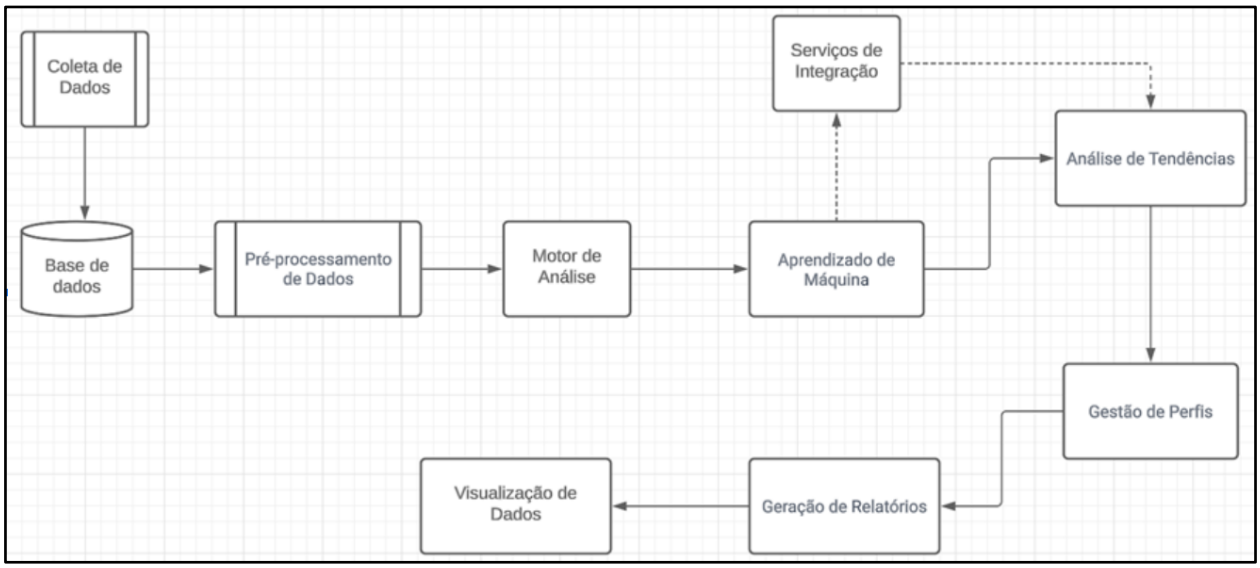
Fase 4 - Apresentação dos Resultados: com o uso da solução automatizada, o gestor terá acesso a uma visualização clara e abrangente dos dados educacionais, permitindo uma compreensão aprofundada das características dos alunos e das tendências nos cursos de ensino superior. Essa compreensão mais precisa dos dados foi fundamental para embasar a tomada de decisão do gestor. Por meio dos gráficos, o gestor poderá visualizar de forma intuitiva informações importantes, como o desempenho acadêmico dos alunos ao longo do tempo, a distribuição das notas, a taxa de aprovação em disciplinas específicas, dentre outros aspectos relevantes para acompanhamento da gestão acadêmica de disciplinas e cursos. Estes resultados fornecerão insights visuais por meio de gráficos que facilitarão a identificação de padrões, tendências ou anomalias, permitindo ao gestor tomar decisões mais embasadas para otimizar a oferta de cursos de ensino superior pela instituição.

Além dos gráficos, os relatórios gerados pela solução automatizada fornecerão uma análise mais detalhada dos dados educacionais. Esses relatórios prestarão suporte ao gestor, que poderá acompanhar aspectos específicos, como a composição do corpo discente, a taxa de evasão estudantil, o desempenho em áreas específicas de conhecimento ou disciplinas, dentre outros indicadores relevantes à gestão acadêmica. Com base nessas informações, o gestor terá uma visão mais completa do cenário acadêmico, permitindo assim a identificação de áreas que requerem melhorias, como a implementação de políticas de retenção de alunos e a adequação da oferta de cursos e disciplinas às demandas de alunos e de mercado. Dessa forma, a solução proposta de mineração de dados educacionais ora estipulado oferece ao gestor a capacidade de realizar análises detalhadas por meio dos gráficos e relatórios disponibilizados. Isso contribuirá diretamente para uma tomada de decisão mais

embasada, permitindo ao gestor otimizar a oferta de cursos de ensino superior pela instituição.

Na Figura 10 são ilustrados os passos de análise de dados, sendo pré-estabelecida cada etapa do processo de análise. A partir da coleta de dados acadêmicos, o tratamento dos dados e a resposta das análises de gráficos e relatórios gerenciais, foi estabelecido um panorama do perfil de alunos de acordo com cursos e disciplinas ofertados pela instituição.

Figura 10: Passos da análise dos dados da coleta à visualização dos resultados



Fonte: Autora (2023)

A Figura 10 ilustra a sequência de etapas necessárias para a criação da solução proposta nesta pesquisa. O processo se inicia com a coleta de dados, realizado idealmente a cada semestre, seguindo os ciclos estabelecidos pela gestão acadêmica ou de acordo com necessidades específicas da instituição. Essa coleta visa obter os dados relevantes para a análise, sendo um ponto crucial para o desenvolvimento da solução.

Após a coleta, os dados são submetidos a um processo de tratamento, no qual a solução realiza a conversão dos dados brutos em metadados adequados para o processamento pelos algoritmos responsáveis pela predição de perfis de aluno. Essa etapa envolve o tratamento de dados ausentes, a normalização, se necessário, e outras técnicas de preparação dos dados para a análise.

Uma vez que os dados estão tratados, a solução utiliza algoritmos de predição de perfil de aluno para gerar métricas e visualizações em gráficos. Essas métricas e visualizações são apresentadas ao gestor, permitindo que ele identifique pontos de melhoria na oferta de cursos com base na compreensão das características e desempenho dos alunos. Essas informações são valiosas para direcionar as ações de marketing da instituição de ensino superior e assim aprimorar a atratividade dos cursos oferecidos para o público-alvo de potenciais candidatos.

3.4 Base de Dados e Plataforma de Ensaios

Neste tópico é apresentada a origem dos dados, juntamente ao processo de coleta e seleção dos atributos selecionados para a realização das análises nos experimentos realizados nesta pesquisa. Dentre os instrumentos para análise de dados foi elaborada uma planilha com os dados recebidos por uma instituição de ensino superior. A base de dados considerada continha dados reais de 459 alunos, segregados em dois cursos de pós-graduação (Governança em TI, com 187 alunos e Data Science, com 272 alunos). Ressalta-se que foram excluídos dados considerados pessoais/sigilosos, tais como nome do aluno, e-mail, RG, CPF, etc.

A Figura 9 apresenta como ficou a base após a retirada dos dados sigilosos. A base refinada ficou com 24 categorias de dados dos alunos, conforme expostas nas Figuras 11 (ingresso do aluno no curso) e 12 (finalização do curso pelo aluno), que exibem o exemplo do curso de pós-graduação de Governança em TI.

Figura 11: Estrutura da base de dados no ingresso do aluno no curso

ALUNO	IDADE	GÊNERO	REGIAO	CURSO	TURMA	EEMN	EEMT	EEME	EGB	EGT	EPG
1	22	M	SP	TRC	A	0	1	0	0	1	0
2	30	M	RJ	TADS	A	0	1	0	0	1	0
3	28	M	BA	TGTI	A	0	1	0	0	1	0
4	24	M	ES	TGTI	A	1	0	0	1	0	0
5	41	M	SP	CC	A	0	1	0	1	0	0
6	27	F	SP	SI	A	0	1	0	1	0	0
7	32	M	MG	CC	A	0	1	0	1	0	0
8	59	M	SP	TRC	A	0	0	1	0	1	0
9	47	M	RJ	TSEG	A	0	0	1	0	1	0
10	29	F	RS	CC	A	1	0	0	1	0	0
11	30	M	MG	TGTI	A	1	0	0	0	1	0
12	25	F	SP	SI	A	1	0	0	1	0	0
13	26	M	PE	TSIN	A	0	1	0	0	1	0
14	22	F	SP	TRC	A	1	0	0	0	1	0
15	28	M	MG	TGTI	A	0	1	0	0	1	0
16	45	M	SP	CC	A	1	0	0	1	0	1
17	27	F	BA	TSIN	A	0	1	0	0	1	0
18	26	M	SP	TGTI	A	1	0	0	0	1	0
19	27	M	CE	CC	A	1	0	0	1	0	0
20	29	F	SP	CC	A	1	0	0	1	0	0
21	56	M	SP	TSIN	A	0	0	1	0	1	0
22	23	M	RJ	TRC	A	0	1	0	0	1	0
23	25	F	SP	TRC	A	0	1	0	0	1	0
24	49	M	SP	TSEG	A	1	0	0	0	1	0
25	23	M	SP	TRC	A	1	0	0	0	1	0
26	30	F	RJ	TGTI	A	1	0	0	0	1	0
27	25	M	SP	TGTI	A	0	1	0	0	1	0
28	30	F	SP	TGTI	A	1	0	0	0	1	1
29	26	M	ES	CC	A	1	0	0	1	0	0
30	29	M	SP	TADS	A	0	1	0	0	1	0

Fonte: Autora (2023)

Na Figura 11 é exibida a situação do aluno no momento de seu ingresso no curso (denominado ‘momento 0’), ou seja, quando o aluno realiza sua matrícula na instituição de ensino. Nesse momento são coletados os dados pessoais de documentos e demais informações sobre o aluno, tais como gênero e idade, dentre outros. Além disso, são registradas informações sobre a formação prévia do aluno no ensino médio e graduação. A seguir são apresentados os atributos relacionados ao momento de matrícula do aluno:

- Aluno – Identificação do aluno na instituição;
- Idade – Idade do aluno;
- Gênero – Identificação do gênero (masculino ou feminino);
- Região – Estado em que o aluno declarou sua residência;
- Curso – Curso que o aluno será matriculado;
- Turma – Turma que o aluno será matriculado;
- EEMN – Escolaridade Ensino Médio Normal
- EEMT – Escolaridade Ensino Médio Técnico
- EEME – Escolaridade Ensino Médio EJA

- EGB – Escolaridade Graduação Bacharel
- EGT – Escolaridade Graduação Tecnólogo
- EPG – Escolaridade Pós-Graduação

Os atributos de perfil de aluno descritos na sua matrícula no curso (‘momento 0’) desempenham papel fundamental para a obtenção dos parâmetros iniciais de perfil do aluno. Ou seja, os atributos são importantes para compreender melhor o aluno, suas características pessoais e sua formação prévia. Tais atributos também serão importantes para o delineamento do perfil do aluno e sua performance ao final do curso. A seguir são apresentados o detalhamento dos atributos do cadastro de aluno quando de sua matrícula no curso.

Aluno: identificação única do aluno na instituição. Cada aluno recebe um número ou código de identificação que o distingue dos demais estudantes;

Idade: idade do aluno calculada a partir de sua data de nascimento. Essa informação é geralmente registrada com o objetivo de fornecer dados demográficos e estatísticas relacionadas à faixa etária dos estudantes;

Gênero: representa a identificação do gênero do aluno, podendo ser masculino ou feminino. Essa informação é importante para fins de análise demográfica, além de auxiliar na promoção de igualdade de gênero e políticas de inclusão;

Região: refere-se ao estado ou região geográfica em que o aluno declarou residir. Essa informação é coletada para análise geográfica, planejamento de recursos e estatísticas relacionadas à distribuição dos alunos em diferentes áreas.

Curso: indica o curso específico em que o aluno será matriculado. Essa informação é fundamental para fins administrativos, acompanhamento acadêmico e elaboração de estatísticas relacionadas à distribuição dos alunos em diferentes cursos;

Turma: representa a turma em que o aluno estará alocado. A turma é uma subdivisão do curso e é geralmente formada por um grupo de alunos que cursa disciplinas em conjunto durante um determinado período letivo;

Escolaridade Ensino Médio Normal (EEMN): indica o nível de escolaridade do aluno referente ao Ensino Médio na modalidade normal, que se trata do curso regular de formação no Ensino Médio;

Escolaridade Ensino Médio Técnico (EEMT): representa o nível de escolaridade do aluno referente ao Ensino Médio na modalidade técnica, caracterizado pela formação profissionalizante, aliando conhecimentos teóricos e práticos em uma área específica;

Escolaridade Ensino Médio EJA (EEME): refere-se ao nível de escolaridade do aluno referente ao Ensino Médio na modalidade de Educação de Jovens e Adultos (EJA). Essa modalidade de ensino é destinada a pessoas que não concluíram seus estudos na idade adequada e desejam obter o diploma do Ensino Médio.

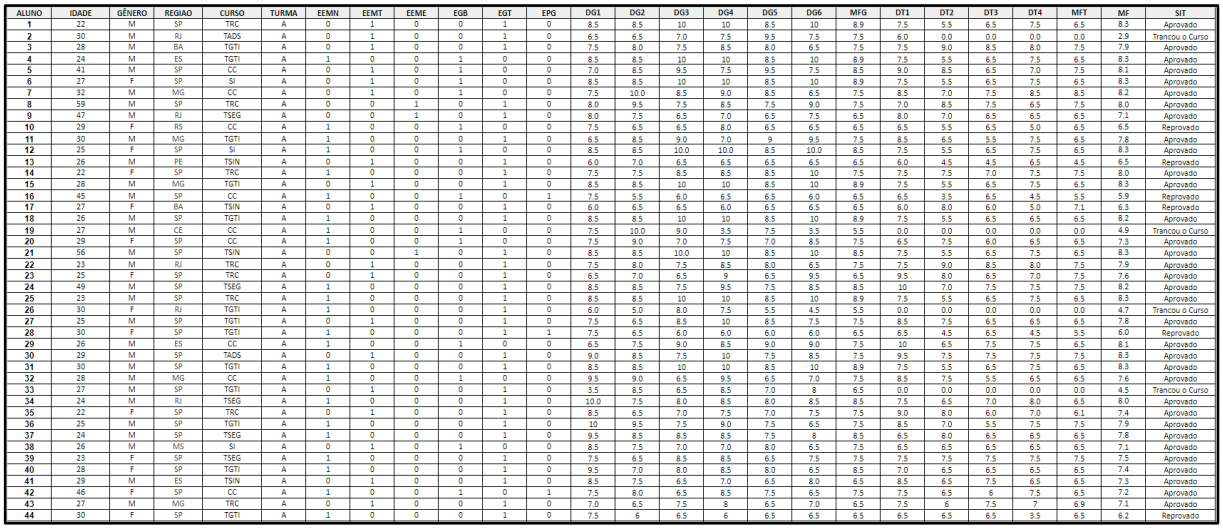
Escolaridade Graduação Bacharel (EGB): indica o nível de escolaridade do aluno referente à graduação na modalidade Bacharelado. O bacharelado é um tipo de curso superior que proporciona uma formação ampla e aprofundada em uma área específica do conhecimento;

Escolaridade Graduação Tecnólogo (EGT): representa o nível de escolaridade do aluno referente à graduação na modalidade Tecnólogo. O tecnólogo é um curso superior de curta duração cujo objetivo é fornecer conhecimentos teóricos e práticos específicos para o exercício de uma profissão;

Escolaridade Pós-Graduação (EPG): refere-se ao nível de escolaridade do aluno referente à pós-graduação, que engloba cursos de especialização.

Após a entrada do aluno e o decorrer do ano letivo são atribuídas notas individuais para cada disciplina cursada por cada aluno no curso em que está matriculado. Ao final do curso, ora denominado ‘’momento 1’, o aluno terá registrado em seu perfil todas as notas e eventos das disciplinas que cursou. As Figuras 12 e 13 expõem a estrutura de dados do aluno ao término do ano letivo.

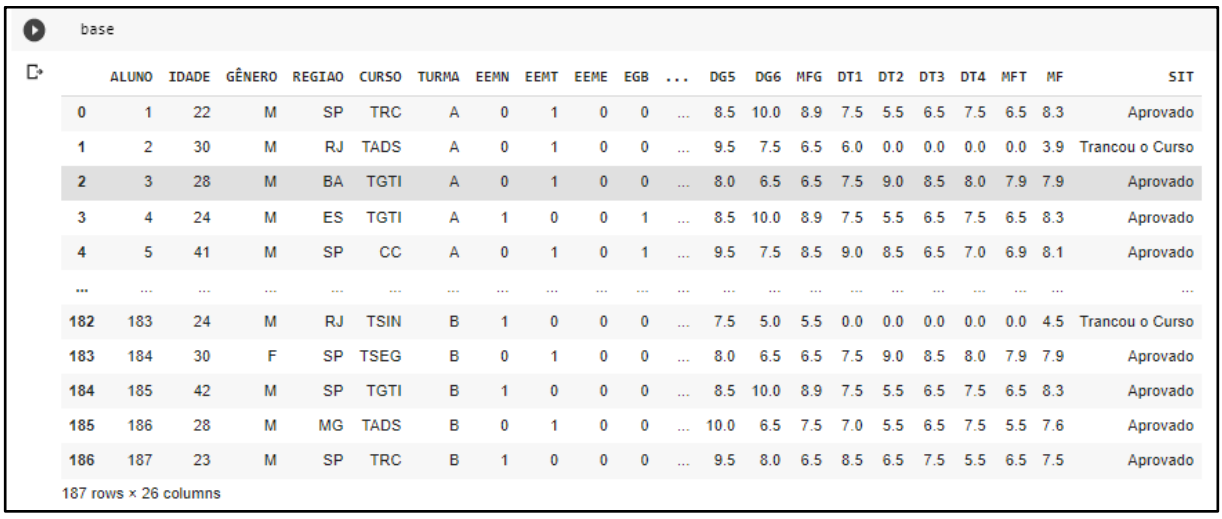
Figura 12: Estrutura da base de dados ao término do ano letivo



ALUNO	IDADE	GÊNERO	REGIAO	CURSO	TURMA	EEMN	EEMT	EEME	EGB	EGT	EPG	DG1	DG2	DG3	DG4	DG5	DG6	MFG	DT1	DT2	DT3	DT4	MFT	MF	SIT
1	22	M	SP	TRC	A	0	1	0	0	1	0	8.5	8.5	10	10	8.5	10	8.9	7.5	5.5	6.5	7.5	6.5	8.3	Aprovado
2	30	M	RJ	TADS	A	0	1	0	0	1	0	6.5	6.5	7.0	7.5	9.5	7.5	7.5	6.0	0.0	0.0	0.0	0.0	2.9	Trancou o Curso
3	28	M	BA	TGTI	A	0	1	0	0	1	0	7.5	8.0	7.5	8.5	8.0	6.5	7.5	7.5	9.0	8.5	8.0	7.5	7.9	Aprovado
4	24	M	ES	TGTI	A	1	0	0	1	0	0	8.5	8.5	10	10	8.5	10	8.9	7.5	5.5	6.5	7.5	6.5	8.3	Aprovado
5	41	M	SP	CC	A	0	1	0	1	0	0	7.0	8.5	9.5	7.5	9.5	7.5	8.5	9.0	8.5	6.5	7.0	7.5	8.1	Aprovado
6	27	F	SP	SI	A	0	1	0	1	0	0	8.5	8.5	10	10	8.5	10	8.9	7.5	5.5	6.5	7.5	6.5	8.3	Aprovado
7	32	M	MG	CC	A	0	1	0	1	0	0	7.5	10.0	8.5	9.0	8.5	6.5	7.5	8.5	7.0	7.5	8.5	8.5	8.2	Aprovado
8	59	M	SP	TRC	A	0	0	1	0	1	0	8.0	9.5	7.5	8.5	7.5	9.0	7.5	7.0	8.5	7.5	6.5	7.5	8.0	Aprovado
9	47	M	RJ	TSEG	A	0	0	1	0	1	0	8.0	7.5	6.5	7.0	6.5	7.5	6.5	8.0	7.0	6.5	6.5	6.5	7.1	Aprovado
10	29	F	RS	CC	A	1	0	0	1	0	0	7.5	6.5	6.5	8.0	6.5	6.5	6.5	6.5	5.5	6.5	5.0	6.5	6.5	Reprovado
11	30	M	MG	TGTI	A	1	0	0	0	1	0	6.5	8.5	9.0	7.0	9	9.5	7.5	8.5	6.5	5.5	7.5	6.5	7.8	Aprovado
12	25	F	SP	SI	A	1	0	0	1	0	0	8.5	8.5	10.0	10.0	8.5	10.0	8.5	7.5	5.5	6.5	7.5	6.5	8.3	Aprovado
13	26	M	PE	TSIN	A	0	1	0	0	1	0	6.0	7.0	6.5	6.5	6.5	6.5	6.5	6.0	4.5	4.5	6.5	4.5	6.5	Reprovado
14	22	F	SP	TRC	A	1	0	0	0	1	0	7.5	7.5	8.5	8.5	8.5	10	7.5	7.5	7.5	7.0	7.5	7.5	8.0	Aprovado
15	28	M	MG	TGTI	A	0	1	0	0	1	0	8.5	8.5	10	10	8.5	10	8.9	7.5	5.5	6.5	7.5	6.5	8.3	Aprovado
16	45	M	SP	CC	A	1	0	0	1	0	1	7.5	5.5	6.0	6.5	6.5	6.0	6.5	6.5	3.5	6.5	4.5	5.5	5.9	Reprovado
17	27	F	BA	TSIN	A	0	1	0	0	1	0	6.0	6.5	6.5	6.0	6.5	6.5	6.5	6.0	8.0	6.0	5.0	7.1	6.3	Reprovado
18	26	M	SP	TGTI	A	1	0	0	0	1	0	8.5	8.5	10	10	8.5	10	8.9	7.5	5.5	6.5	6.5	6.5	8.2	Aprovado
19	27	M	CE	CC	A	1	0	0	1	0	0	7.5	10.0	9.0	3.5	7.5	3.5	5.5	0.0	0.0	0.0	0.0	0.0	4.9	Trancou o Curso
20	29	F	SP	CC	A	1	0	0	1	0	0	7.5	9.0	7.0	7.5	7.0	8.5	7.5	6.5	7.5	6.0	6.5	6.5	7.3	Aprovado
21	56	M	SP	TSIN	A	0	0	1	0	1	0	8.5	8.5	10.0	10	8.5	10	8.5	7.5	5.5	6.5	7.5	6.5	8.3	Aprovado
22	23	M	RJ	TRC	A	0	1	0	0	1	0	7.5	8.0	7.5	8.5	8.0	6.5	7.5	7.5	9.0	8.5	8.0	7.5	7.9	Aprovado
23	25	F	SP	TRC	A	0	1	0	0	1	0	6.5	7.0	6.5	9	6.5	9.5	6.5	9.5	8.0	6.5	7.0	7.5	7.6	Aprovado
24	49	M	SP	TSEG	A	1	0	0	0	1	0	8.5	8.5	7.5	9.5	7.5	8.5	8.5	10	7.0	7.5	7.5	7.5	8.2	Aprovado
25	23	M	SP	TRC	A	1	0	0	0	1	0	8.5	8.5	10	10	8.5	10	8.9	7.5	5.5	6.5	7.5	6.5	8.3	Aprovado
26	30	F	RJ	TGTI	A	1	0	0	0	1	0	6.0	5.0	8.0	7.5	5.5	4.5	5.5	0.0	0.0	0.0	0.0	0.0	4.7	Trancou o Curso
27	25	M	SP	TGTI	A	0	1	0	0	1	0	7.5	6.5	8.5	10	8.5	7.5	7.5	8.5	7.5	6.5	6.5	6.5	7.8	Aprovado
28	30	F	SP	TGTI	A	1	0	0	0	1	1	7.5	6.5	6.0	6.0	6.0	6.0	6.5	6.5	4.5	6.5	4.5	5.5	6.0	Reprovado
29	26	M	ES	CC	A	1	0	0	1	0	0	6.5	7.5	9.0	8.5	9.0	9.0	7.5	10	6.5	7.5	7.5	6.5	8.1	Aprovado
30	29	M	SP	TADS	A	0	1	0	0	1	0	9.0	8.5	7.5	10	7.5	8.5	7.5	9.5	7.5	7.5	7.5	7.5	8.3	Aprovado
31	30	M	SP	TGTI	A	1	0	0	0	1	0	8.5	8.5	10	10	8.5	10	8.9	7.5	5.5	6.5	7.5	6.5	8.3	Aprovado
32	28	M	MG	CC	A	1	0	0	1	0	0	9.5	9.0	6.5	9.5	6.5	7.0	7.5	8.5	7.5	5.5	6.5	6.5	7.6	Aprovado
33	27	M	SP	TGTI	A	0	1	0	0	1	0	3.5	8.5	6.5	8.5	7.0	8	6.5	0.0	0.0	0.0	0.0	0.0	4.5	Trancou o Curso
34	24	M	RJ	TSEG	A	1	0	0	0	1	0	10.0	7.5	8.0	8.5	8.0	8.5	8.5	7.5	6.5	7.0	8.0	6.5	8.0	Aprovado
35	22	F	SP	TRC	A	0	1	0	0	1	0	8.5	6.5	7.0	7.5	7.0	7.5	7.5	9.0	8.0	6.0	7.0	6.1	7.4	Aprovado
36	25	M	SP	TGTI	A	1	0	0	0	1	0	10	9.5	7.5	9.0	7.5	6.5	7.5	8.5	7.0	5.5	7.5	7.5	7.9	Aprovado
37	24	M	SP	TSEG	A	1	0	0	0	1	0	9.5	8.5	8.5	8.5	7.5	8	8.5	6.5	8.0	6.5	6.5	6.5	7.8	Aprovado
38	26	M	MS	SI	A	0	1	0	1	0	0	8.5	7.5	7.0	7.0	8.0	6.5	7.5	6.5	6.5	6.5	6.5	6.5	7.1	Aprovado
39	23	F	SP	TSEG	A	1	0	0	0	1	0	7.5	6.5	8.5	8.5	6.5	7.5	7.5	7.5	7.5	7.5	7.5	7.5	7.5	Aprovado
40	28	F	SP	TGTI	A	1	0	0	0	1	0	9.5	7.0	8.0	8.5	8.0	6.5	8.5	7.0	6.5	6.5	6.5	6.5	7.4	Aprovado
41	29	M	ES	TSIN	A	0	1	0	0	1	0	8.5	7.5	6.5	7.0	6.5	8.0	6.5	8.5	6.5	7.5	6.5	6.5	7.3	Aprovado
42	46	F	SP	CC	A	1	0	0	1	0	1	7.5	8.0	6.5	8.5	7.5	6.5	7.5	7.5	6.5	6	7.5	6.5	7.2	Aprovado
43	27	M	MG	TRC	A	0	1	0	0	1	0	7.0	6.5	7.5	8	6.5	7.0	6.5	7.5	6	7.5	7	6.9	7.1	Aprovado
44	30	F	SP	TGTI	A	1	0	0	0	1	0	7.5	6	6.5	6	6.5	6.5	6.5	6.5	6.5	6.5	3.5	6.5	6.2	Reprovado

Fonte: Autora (2023)

Figura 13: Estrutura da base de dados ao término do ano letivo no Google Colab



base

	ALUNO	IDADE	GÊNERO	REGIAO	CURSO	TURMA	EEMN	EEMT	EEME	EGB	...	DG5	DG6	MFG	DT1	DT2	DT3	DT4	MFT	MF	SIT
0	1	22	M	SP	TRC	A	0	1	0	0	...	8.5	10.0	8.9	7.5	5.5	6.5	7.5	6.5	8.3	Aprovado
1	2	30	M	RJ	TADS	A	0	1	0	0	...	9.5	7.5	6.5	6.0	0.0	0.0	0.0	0.0	3.9	Trancou o Curso
2	3	28	M	BA	TGTI	A	0	1	0	0	...	8.0	6.5	6.5	7.5	9.0	8.5	8.0	7.9	7.9	Aprovado
3	4	24	M	ES	TGTI	A	1	0	0	1	...	8.5	10.0	8.9	7.5	5.5	6.5	7.5	6.5	8.3	Aprovado
4	5	41	M	SP	CC	A	0	1	0	1	...	9.5	7.5	8.5	9.0	8.5	6.5	7.0	6.9	8.1	Aprovado
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
182	183	24	M	RJ	TSIN	B	1	0	0	0	...	7.5	5.0	5.5	0.0	0.0	0.0	0.0	0.0	4.5	Trancou o Curso
183	184	30	F	SP	TSEG	B	0	1	0	0	...	8.0	6.5	6.5	7.5	9.0	8.5	8.0	7.9	7.9	Aprovado
184	185	42	M	SP	TGTI	B	1	0	0	0	...	8.5	10.0	8.9	7.5	5.5	6.5	7.5	6.5	8.3	Aprovado
185	186	28	M	MG	TADS	B	0	1	0	0	...	10.0	6.5	7.5	7.0	5.5	6.5	7.5	5.5	7.6	Aprovado
186	187	23	M	SP	TRC	B	1	0	0	0	...	9.5	8.0	6.5	8.5	6.5	7.5	5.5	6.5	7.5	Aprovado

187 rows × 26 columns

Fonte: Autora (2023)

Nas Figuras 12 e 13 são apresentadas as categorias de dados dispostas na base de dados ao término do ano letivo do aluno, conforme descrição a seguir:

- Aluno – Identificação do aluno na instituição;
- Idade – Idade do aluno;
- Gênero – Identificação do gênero (masculino ou feminino);
- Região – Estado em que o aluno declarou sua residência;
- Curso – Curso que o aluno está/será matriculado;
- Turma – Turma que o aluno está/será matriculado;

- EEMN – Escolaridade Ensino Médio Normal
- EEMT – Escolaridade Ensino Médio Técnico
- EEME – Escolaridade Ensino Médio EJA
- EGB – Escolaridade Graduação Bacharel
- EGT – Escolaridade Graduação Tecnólogo
- EPG – Escolaridade Pós-Graduação
- DG1 – Nota da disciplina Gerencial 1
- DG2 – Nota da disciplina Gerencial 2
- DG3 – Nota da disciplina Gerencial 3
- DG4 – Nota da disciplina Gerencial 4
- DG5 – Nota da disciplina Gerencial 5
- DG6 – Nota da disciplina Gerencial 6
- MFG – Média final das disciplinas gerenciais
- DT1 – Nota da disciplina Técnica 1
- DT2 – Nota da disciplina Técnica 2
- DT3 – Nota da disciplina Técnica 3
- DT4 – Nota da disciplina Técnica 4
- MFT – Média final das disciplinas técnicas
- MF – Média final total das disciplinas (gerenciais e técnicas)
- SIT – Situação (aprovado, reprovado ou trancou o curso)

Após o tratamento preliminar dos dados da base selecionada para aplicação nesta pesquisa foram selecionadas as seguintes categorias: idade, gênero, região, curso, turma, escolaridade (ensino normal, técnico ou EJA), graduação (bacharel ou tecnólogo) e pós-graduação, disciplinas gerenciais (DG), disciplinas técnicas (DT), média final das disciplinas gerenciais (MFG), média final das disciplinas técnicas (MFT), média final total das disciplinas (MF) e situação (SIT: aprovado, reprovado ou Trancou o curso).

3.5 Transformação e Preparação dos Dados

Antes de iniciar as análises foi realizado o pré-processamento dos dados da base selecionada para a adequação do idioma e origem, bem como a identificação e

exclusão de dados nulos ou incompletos e colunas, através dos comandos: dataset.isnull().sum() e dataset = dataset.drop(columns = []) do Google Colab. Dessa forma, foi possível manusear a base de dados para considerar apenas as colunas que possuem os atributos e relacionamentos considerados para a análise desta pesquisa, conforme indicado na Figura 14.

Figura 14: Estrutura da base de dados após exclusão

```
dataset
```

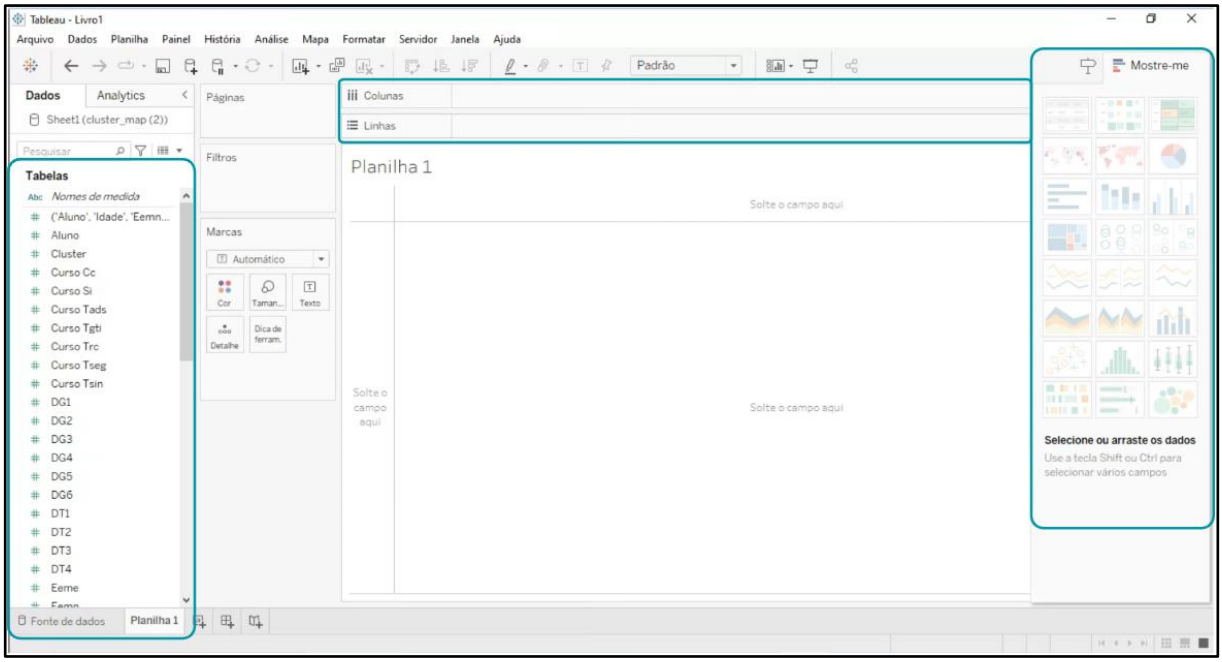
	ALUNO	IDADE	GÊNERO	REGIAO	CURSO	TURMA	EEMN	EEMT	EEME	EGB	...	DG5	DG6	MFG	DT1	DT2	DT3	DT4	MFT	MF	SIT
0	1	22	M	SP	TRC	A	0	1	0	0	...	8.5	10.0	8.9	7.5	5.5	6.5	7.5	6.5	8.3	Aprovado
1	2	30	M	RJ	TADS	A	0	1	0	0	...	9.5	7.5	6.5	6.0	0.0	0.0	0.0	0.0	3.9	Trancou_o_Curso
2	3	28	M	BA	TGTI	A	0	1	0	0	...	8.0	6.5	6.5	7.5	9.0	8.5	8.0	7.9	7.9	Aprovado
3	4	24	M	ES	TGTI	A	1	0	0	1	...	8.5	10.0	8.9	7.5	5.5	6.5	7.5	6.5	8.3	Aprovado
4	5	41	M	SP	CC	A	0	1	0	1	...	9.5	7.5	8.5	9.0	8.5	6.5	7.0	6.9	8.1	Aprovado
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
182	183	24	M	RJ	TSIN	B	1	0	0	0	...	7.5	5.0	5.5	0.0	0.0	0.0	0.0	0.0	4.5	Trancou_o_Curso
183	184	30	F	SP	TSEG	B	0	1	0	0	...	8.0	6.5	6.5	7.5	9.0	8.5	8.0	7.9	7.9	Aprovado
184	185	42	M	SP	TGTI	B	1	0	0	0	...	8.5	10.0	8.9	7.5	5.5	6.5	7.5	6.5	8.3	Aprovado
185	186	28	M	MG	TADS	B	0	1	0	0	...	10.0	6.5	7.5	7.0	5.5	6.5	7.5	5.5	7.6	Aprovado
186	187	23	M	SP	TRC	B	1	0	0	0	...	9.5	8.0	6.5	8.5	6.5	7.5	5.5	6.5	7.5	Aprovado

187 rows × 26 columns

Fonte: Autora (2023)

Após esta etapa de limpeza dos dados foi realizado o processo de análise dos relacionamentos entre os atributos (colunas) da base de dados. Para as primeiras análises dos dados desta base pré-processada foi utilizado o software *Tableau* para análise de dados. Como resultado, pôde-se ter uma visão ampla das relações entre os atributos eleitos como mais importantes como, por exemplo, resultado (aprovado ou reprovado) por gênero, por escolaridade, por idade e pelos créditos conquistados pelo aluno. Tais relações expressam o desempenho dos alunos segregado por determinados atributos (gênero, escolaridade, idade e créditos). Na Figura 15 é exposta à tela do software *Tableau.*

Figura 15: Software de Mineração de dados Educacionais - *Tableau*



Fonte: Autora (2023)

O *Tableau* é uma ferramenta interativa e de fácil utilização para a visualização de dados. Ele permite que os usuários importem dados de várias fontes, como arquivos CSV, Excel e bancos de dados, dentre outros, e assim criem visualizações interativas, painéis e relatórios. Com esta ferramenta foram realizados os cruzamentos entre os atributos mencionados no capítulo anterior (3.3), conforme também indicado no Quadro 4.

Quadro 4: Cruzamento de atributos para análises - *Tableau*

Agrupamento	Atributos
1	Gênero x Curso de Graduação x Situação
2	Gênero x Curso de Graduação x Idade
3	Pós-graduação x Idades
4	Curso de Pós-graduação x Idade x Região x Curso de graduação
5	EEMN x Médias das disciplinas
6	EEMT x Médias das disciplinas
7	EEME x Médias das disciplinas
8	Aprovados x Curso de Graduação
9	Reprovados ou Trancou o curso x Curso de Graduação
10	Aprovados, Curso de graduação TGTI e Idades

Fonte: Autora (2023)

Foram realizados cruzamentos entre diferentes atributos utilizando-se a ferramenta de mineração de dados *Tableau*. Os cruzamentos foram feitos da seguinte forma: 1 - Cruzamento entre Gênero, Curso de Graduação e Situação; 2 - Cruzamento entre Gênero, Curso de Graduação e Idade; 3 - Cruzamento entre Pós-graduação e Idade; 4 - Cruzamento entre Curso de Pós-graduação, Idade, Região e Curso de Graduação; 5 - Cruzamento entre Escolaridade Ensino Médio Normal (EEMN) e Médias das disciplinas; 6 - Cruzamento entre Escolaridade Ensino Médio Técnico (EEMT) e Médias das disciplinas; 7 - Cruzamento entre Escolaridade Ensino Médio EJA (EEME) e Médias das disciplinas; 8 - Cruzamento entre Aprovados e Curso de Graduação; 9 - Cruzamento entre Reprovados ou Trancou o curso e Curso de Graduação e, por fim; 10 - Cruzamento entre Aprovados, Curso de Graduação TGTI e Idade.

Estes cruzamentos entre diferentes atributos foram realizados com o objetivo de descobrir o perfil dos alunos. Eles permitiram analisar as relações entre os atributos mencionados e obter percepções relevantes sobre os dados educacionais em questão. Por meio destas análises foi possível identificar padrões e relações entre o gênero dos alunos, o curso de graduação, a situação acadêmica, a idade, a

escolaridade, as médias das disciplinas e outras informações relevantes. A visão destes padrões e relações forneceram uma compreensão mais aprofundada do perfil dos alunos, possibilitando melhorar a tomada de decisão do gestor educacional, que passa a se embasar em padrões para assim promover o desenvolvimento de estratégias educacionais mais direcionadas.

Ainda no processo de pré-processamento e análise dos dados considerados para este estudo foi utilizado também a ferramenta exploratória o *Google Colab,* com uso da linguagem *Python* para a criação de um notebook para aprofundar a análise das relações de atributos e conduzir alguns experimentos com a base de dados estabelecida. As bibliotecas utilizadas para esta análise com Python foram:

Pandas - Análise e manipulação de dados;

Numpy - Uso de funções para se trabalhar com computação numérica;

Matplotlib - Visualização de dados e plotagem gráfica;

Sklearn - Aplicação prática de machine learning;

Seaborn - Visualização estatística de dados.

Das bibliotecas de Python acima mencionadas, a biblioteca *Pandas* (import pandas as pd) é uma biblioteca responsável pela maior parte de análise de dados. Seu nome é derivado do termo ‘dados de painel’ (*panel data*), um termo econométrico utilizado para se referir a conjuntos de dados estruturados multidimensionais. Ela possui código aberto e uso gratuito (sob uma licença BSD).

NumPy (**import** numpy **as** np) é o pacote fundamental para computação científica em Python. É uma biblioteca Python que fornece um objeto *array* multidimensional, vários objetos derivados (como *arrays* e matrizes mascarados) e uma variedade de rotinas para operações rápidas em *arrays*, incluindo matemática, lógica, manipulação de formas, classificação, seleção, E/S, transformadas discretas de Fourier, álgebra linear básica, operações estatísticas básicas, simulação aleatória e muito mais (NUMPY.ORG, 2022).

Matplotlib (import matplotlib.pyplot as plt) é uma biblioteca abrangente para criar visualizações estáticas, animadas e interativas em Python. Um grande número de

pacotes de terceiros estende e se baseia na funcionalidade Matplotlib, incluindo várias interfaces de plotagem de nível superior (seaborn, HoloViews, ggplot, etc.) e um kit de ferramentas de projeção e mapeamento (Cartopy) (MATPLOTLIB.ORG, 2022).

Scikit-learn (from sklearn import linear_model) é uma ferramenta simples e eficiente para análise preditiva de dados, acessível e eficiente em vários contextos, construído em NumPy, SciPy e matplotlib de código aberto (SCIKIT-LEARN.ORG, 2022).

Seaborn (import seaborn as sns) é uma biblioteca para elaboração de gráficos estatísticos em Python. Ela se baseia no matplotlib e se integra intimamente às estruturas de dados do Pandas. A Seaborn ajuda você a explorar e entender seus dados. Suas funções de plotagem operam em *dataframes* e *arrays* contendo conjuntos de dados inteiros e realizam internamente o mapeamento semântico e a agregação estatística necessários para produzir gráficos informativos (SEABORN.PYDATA.ORG, 2022).

Para uso do *Google Colab* foi realizada a discretização dos dados considerados nesta pesquisa, de modo a prepará-los para as análises a serem realizadas. A discretização de dados é um dos processos mais utilizados no pré-processamento dos dados, visando padronizá-los para sua utilização durante o processo de mineração de dados (NOETZOLD; PERTILE. 2021). A discretização de dados é responsável por preparar o banco utilizado, de modo que resulte em resultados corretos, sem ruídos ou *outliers* que prejudicariam os julgamentos mediante os resultados obtidos (LIU *et al.*, 2002; GUANDALINE, 2016). Para tanto, as variáveis categóricas foram transformadas em variáveis contínuas, de modo a possibilitar o manuseio e exploração dos atributos no *Google Colab.* O processo de discretização dos dados é exposto nas Figuras 16 e 17.

Figura 16: Discretização de dados – Gênero

```
##Substituindo a variável sexo para 0 (F) e 1 (M)
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
dataset['GENERO'] = le.fit_transform(dataset['GÊNERO'])
dataset.head(10)
```

	ALUNO	IDADE	GÊNERO	REGIAO	CURSO	TURMA	EEMN	EEMT	EEME	EGB	...	DG6	MFG	DT1	DT2	DT3	DT4	MFT	MF	SIT	GENERO
0	1	22	M	SP	TRC	A	0	1	0	0	...	10.0	8.9	7.5	5.5	6.5	7.5	6.5	8.3	Aprovado	1
1	2	30	M	RJ	TADS	A	0	1	0	0	...	7.5	6.5	6.0	0.0	0.0	0.0	0.0	3.9	Trancou_o_Curso	1
2	3	28	M	BA	TGTI	A	0	1	0	0	...	6.5	6.5	7.5	9.0	8.5	8.0	7.9	7.9	Aprovado	1
3	4	24	M	ES	TGTI	A	1	0	0	1	...	10.0	8.9	7.5	5.5	6.5	7.5	6.5	8.3	Aprovado	1
4	5	41	M	SP	CC	A	0	1	0	1	...	7.5	8.5	9.0	8.5	6.5	7.0	6.9	8.1	Aprovado	1
5	6	27	F	SP	SI	A	0	1	0	1	...	10.0	8.9	7.5	5.5	6.5	7.5	6.5	8.3	Aprovado	0
6	7	32	M	MG	CC	A	0	1	0	1	...	6.5	6.5	8.5	7.0	7.5	8.5	7.5	8.2	Aprovado	1
7	8	59	M	SP	TRC	A	0	0	1	0	...	9.0	7.5	7.0	8.5	7.5	6.5	6.5	8.0	Aprovado	1
8	9	47	M	RJ	TSEG	A	0	0	1	0	...	7.5	6.5	8.0	7.0	6.5	6.5	6.9	7.1	Aprovado	1
9	10	29	F	RS	CC	A	1	0	0	1	...	6.5	6.5	6.5	5.5	6.5	5.0	5.5	6.5	Reprovado	0

10 rows × 27 columns

Fonte: Autora (2023)

Na Figura 16 na categoria Gênero foi aplicado o comando “.fit_transform” para transformar os valores categóricos “Masculino” e “Feminino” em valores numéricos 0 e 1, criando uma nova coluna a direta da base, sendo que “Masculino” assume o valor “1” e “Feminino” o valor “0”. Após a transformação foi utilizado o comando “dataset = dataset.drop(columns = ['GÊNERO']”, para excluir a coluna original com valores categóricos como apresentado na Figura 17.

Figura 17: Discretização de dados - Região, Curso, Turma e 'SIT'- Situação

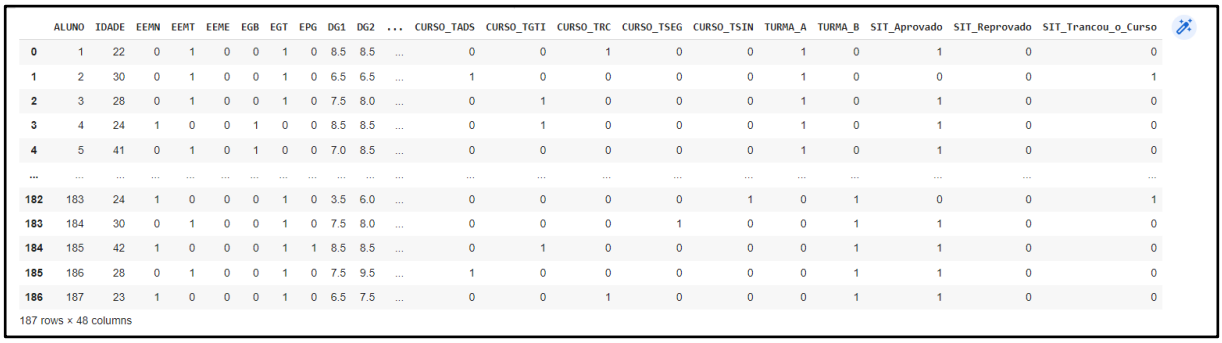
```
dataset = dataset.drop(columns = ['GÊNERO'])

dataset = pd.get_dummies (data = dataset, columns = ['REGIAO','CURSO','TURMA','SIT'] )
dataset
```

Fonte: Autora (2023)

Na Figura 17 além da exclusão da coluna [GÊNERO], para as categorias “REGIAO','CURSO','TURMA','SIT” foi utilizado o comando “pd.get_dummies”. Este comando é uma função da biblioteca Pandas usada para criar variáveis *dummy* (também conhecidas como variáveis indicadoras) a partir de uma variável categórica. Variáveis *dummy* são variáveis binárias que indicam a presença ou ausência de uma determinada categoria em uma variável categórica. Essas variáveis são úteis, pois ao lidar com algoritmos de aprendizado de máquina que requerem entradas numéricas,

ela recebe como entrada uma coluna ou um DataFrame, contendo variáveis categóricas ‘texto ou *strings*’. Ela retorna um DataFrame expandido, no qual cada categoria da variável categórica original é transformada em uma nova coluna binária (dummy). Essas novas colunas indicam a presença ou ausência da categoria em cada observação, ou seja, variáveis binárias (dummies). Na Figura 18 é exposta a tabela após as ações de discretização das categorias da base de dados.

Figura 18: Discretização de dados - Categorias: Região, Curso, Turma e 'SIT'- Situação



	ALUNO	IDADE	EEMN	EEMT	EEME	EGB	EGT	EPG	DG1	DG2	...	CURSO_TADS	CURSO_TGTI	CURSO_TRC	CURSO_TSEG	CURSO_TSIN	TURMA_A	TURMA_B	SIT_Aprovado	SIT_Reprovado	SIT_Trancou_o_Curso
0	1	22	0	1	0	0	1	0	8.5	8.5	...	0	0	1	0	0	1	0	1	0	0
1	2	30	0	1	0	0	1	0	6.5	6.5	...	1	0	0	0	0	1	0	0	0	1
2	3	28	0	1	0	0	1	0	7.5	8.0	...	0	1	0	0	0	1	0	1	0	0
3	4	24	1	0	0	1	0	0	8.5	8.5	...	0	1	0	0	0	1	0	1	0	0
4	5	41	0	1	0	1	0	0	7.0	8.5	...	0	0	0	0	0	1	0	1	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
182	183	24	1	0	0	0	1	0	3.5	6.0	...	0	0	0	0	1	0	1	0	0	1
183	184	30	0	1	0	0	1	0	7.5	8.0	...	0	0	0	1	0	0	1	1	0	0
184	185	42	1	0	0	0	1	1	8.5	8.5	...	0	1	0	0	0	0	1	1	0	0
185	186	28	0	1	0	0	1	0	7.5	9.5	...	1	0	0	0	0	0	1	1	0	0
186	187	23	1	0	0	0	1	0	6.5	7.5	...	0	0	1	0	0	0	1	1	0	0

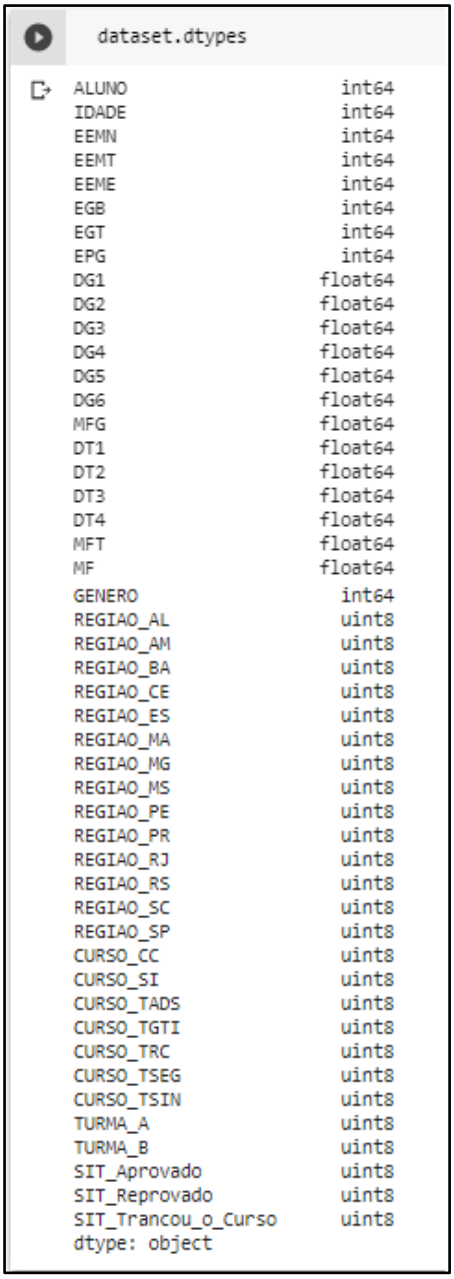
187 rows × 48 columns

Fonte: Autora (2023)

Na Figura 18 foi apresentado resultado após o comando “*pd.get_dummies*” para transformar os dados dos itens expressados em textos por representações binárias, em categorias (0 e 1), considerando-se “0” para ‘não’ e “1” para ‘sim’, de modo a preparar essas categorias de dados para análises mais eficientes da trajetória escolar durante o curso e da formação superior dos alunos.

Agora com todas as colunas devidamente tratadas, são ajustados os dados com comando [dataset.dtypes], com o qual é possível observar o tipo dos dados de nosso dataset, como mostra na Figura 19.

Figura 19: Dataset e visualização de tipos de variáveis



```
dataset.dtypes

ALUNO                 int64
IDADE                 int64
EEMN                  int64
EEMT                  int64
EEME                  int64
EGB                   int64
EGT                   int64
EPG                   int64
DG1                 float64
DG2                 float64
DG3                 float64
DG4                 float64
DG5                 float64
DG6                 float64
MFG                 float64
DT1                 float64
DT2                 float64
DT3                 float64
DT4                 float64
MFT                 float64
MF                  float64
GENERO                int64
REGIAO_AL             uint8
REGIAO_AM             uint8
REGIAO_BA             uint8
REGIAO_CE             uint8
REGIAO_ES             uint8
REGIAO_MA             uint8
REGIAO_MG             uint8
REGIAO_MS             uint8
REGIAO_PE             uint8
REGIAO_PR             uint8
REGIAO_RJ             uint8
REGIAO_RS             uint8
REGIAO_SC             uint8
REGIAO_SP             uint8
CURSO_CC              uint8
CURSO_SI              uint8
CURSO_TADS            uint8
CURSO_TGTI            uint8
CURSO_TRC             uint8
CURSO_TSEG            uint8
CURSO_TSIN            uint8
TURMA_A               uint8
TURMA_B               uint8
SIT_Aprovado          uint8
SIT_Reprovado         uint8
SIT_Trancou_o_Curso   uint8
dtype: object
```

Fonte: Autora (2023)

Na Figura 19 observa-se três tipos de variáveis (int64, uint8 e float64). Para aplicação do algoritmo foi necessário transformar essas variáveis em apenas um tipo float64 *(*um tipo de dado numérico de ponto flutuante com 64 bits de precisão. Números decimais com casas decimais, permitindo tanto valores inteiros como fracionários*).* Isto ajuda o algoritmo na precisão numérica e na conformidade com requisitos de algoritmos. Alguns algoritmos de aprendizado de máquina e estatística

têm requisitos específicos em relação aos tipos de dados que podem ser usados como entrada.

Desta forma, foi aplicado o comando [.astype(dtype='float64')], estamos convertendo o tipo de dado da coluna para float64, ou seja, para números de ponto flutuante de 64 bits. Em sequência, realizou-se a redução de dimensionalidade com comando pca = PCA(n_components = 2).fit_transform(dataset), reduzindo assim a dimensionalidade para um espaço bidimensional. Isso permitirá a visualização dos dados em um gráfico de dispersão em duas dimensões. Nesta fase foram realizados testes de algoritmos pela solução delineada para prever um perfil de aluno mediante as categorias analisadas.

3.5.1 Aplicação do Algoritmo K-Means

Neste tópico são apresentados os passos para realização do uso do algoritmo de clusterização K-Means. Sua aplicação consiste em um processo de clusterização de dados, onde o objetivo é agrupar os elementos em clusters de acordo com suas características similares, conforme exposto na Figura 20.

Figura 20: Aplicação do Algoritmo K-Means

```
# Aplica redução de dimensionalidade
# Transforma as variáveis em 2 variaveis principais. Esse método utiliza Algebra Linear pra identificar semelhança
# entre os dados e assim "juntar" as variaveis, medindo a semelhança pela variância.
pca = PCA(n_components = 2).fit_transform(dataset)

# Determinando um range de K
k_range = range(1,10)

# Aplicando o modelo K-Means para cada valor de K
k_means_var = [KMeans(n_clusters = k).fit(pca) for k in k_range]

# Ajustando o centróide do cluster para cada modelo
centroids = [X.cluster_centers_ for X in k_means_var]

# Calculando a distância euclidiana de cada ponto de dado para o centróide
k_euclid = [cdist(pca, cent, 'euclidean') for cent in centroids]
dist = [np.min(ke, axis = 1) for ke in k_euclid]
```

Fonte: Autora (2023)

Após a redução de dimensionalidade, a Figura 20 apresenta o ‘K_range’ para determinar um número de *clusters*, a abordagem permite testar diferentes valores de k para encontrar o número ótimo de *clusters* para o conjunto de dados em análise.

Nesse caso, ‘*k_range’* está variando entre 1 a 10. Depois aplica-se o modelo K-Means para cada valor de K determinado. Na sequência ajusta-se o *centróide* do cluster para cada modelo. Uma vez feita a aplicação do algoritmo, calcula-se a distância euclidiana entre os pontos do conjunto de dados *‘pca’* e os *‘centróides’* de cada modelo de *‘K-means’* na lista *‘centroids’*.

A distância euclidiana é uma medida comumente utilizada para calcular a distância entre dois pontos em um espaço Euclidiano. Ela é baseada no Teorema de Pitágoras e mede a distância "em linha reta" entre dois pontos em um espaço de coordenadas. Ela é amplamente utilizada em várias áreas, incluindo ciência de dados, aprendizado de máquina e reconhecimento de padrões. Ela é especialmente útil em algoritmos de clusterização, como o *‘K-means’*, onde a ideia é agrupar pontos com base em sua proximidade. Ao calcular a distância euclidiana entre os pontos e os *centróides*, podemos quantificar a distância ou similaridade entre eles. Isso nos permite identificar quais pontos estão mais próximos de um determinado *centróide* e, portanto, provavelmente pertencem a um determinado cluster (FAISAL, *et al.*, 2020).

Fórmula da Distância Euclidiana: $\mathbf{d=\sqrt{\sum_{i=1}^{n}(a_i-b_i)^2}}$ **(1)**

Após o cálculo da Distância Euclidiana foi realizado o cálculo da soma dos quadrados para avaliar a qualidade dos *clusters* gerados pelo algoritmo utilizado nesta clusterização, conforme exposto na Figura 21.

Figura 21: Aplicação do modelo de Clusterização - ‘Somas - intra_cluster e total’

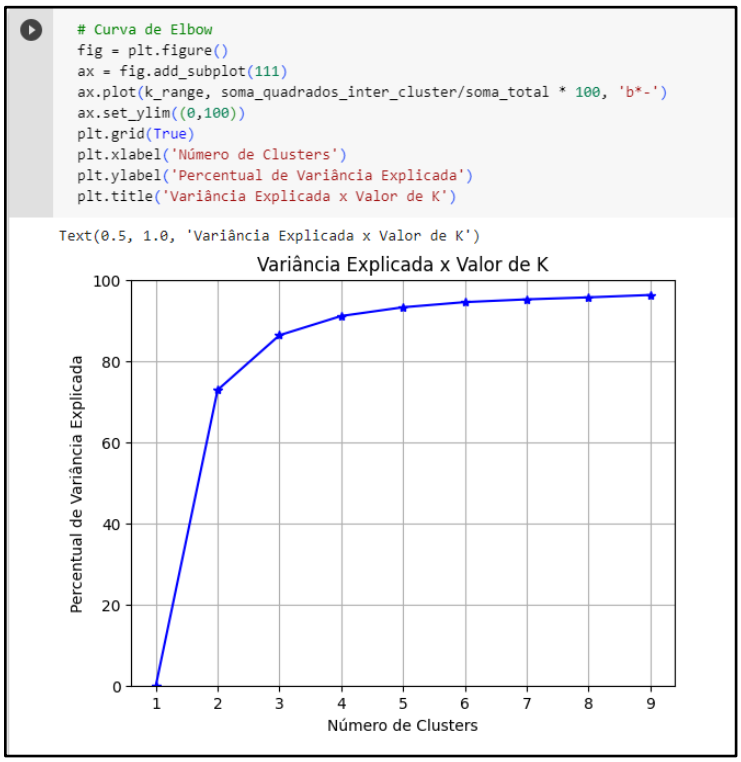
```
# Soma dos quadrados das distâncias dentro do cluster
soma_quadrados_intra_cluster = [sum(d**2) for d in dist]

# Soma total dos quadrados
soma_total = sum(pdist(pca)**2)/pca.shape[0]

# Soma dos quadrados entre clusters
soma_quadrados_inter_cluster = soma_total - soma_quadrados_intra_cluster
```

Fonte: Autora (2023)

A soma dos quadrados das distâncias dentro do cluster ‘intra_cluster’, também conhecida como SSE (*Sum of Squared Errors*) é uma métrica utilizada para avaliar a qualidade dos clusters gerados por algoritmos de clusterização. Ela representa a soma dos quadrados das distâncias entre cada ponto e o centróide do *cluster* ao qual pertence.

Após esta medida foi calculada a ‘soma total dos quadrados’, onde o *‘pdist(pca)’* é responsável por calcular as distâncias entre todos os pares de pontos no conjunto de dados transformado pelo PCA. A função *‘pdist’* retorna uma matriz de distâncias condensada, que contém todas as distâncias únicas entre os pontos. Essa métrica é tradicionalmente utilizada para avaliar a qualidade do agrupamento. Quanto menor for o valor, mais compactos e bem definidos são os clusters. E por último: ‘soma_quadrados_inter_cluster = soma_total - soma_quadrados_intra_cluster’ Essa métrica é usada para avaliar a separação entre os clusters. Quanto maior for o valor de ‘soma_quadrados_inter_cluster’, maior será a separação entre os clusters e melhor será a qualidade do agrupamento, conforme exposto na Figura 22.

Figura 22: Curva de Elbow (cotovelo)




```
# Curva de Elbow
fig = plt.figure()
ax = fig.add_subplot(111)
ax.plot(k_range, soma_quadrados_inter_cluster/soma_total * 100, 'b*-')
ax.set_ylim((0,100))
plt.grid(True)
plt.xlabel('Número de Clusters')
plt.ylabel('Percentual de Variância Explicada')
plt.title('Variância Explicada x Valor de K')
```

Fonte: Autora (2023)

A partir destes primeiros cálculos é possível visualizar a ‘curva de Elbow’ (também conhecida por ‘cotovelo’), que é um gráfico que mostra a relação entre o número de *clusters* e a soma dos quadrados das distâncias dentro do *cluster*. É uma ferramenta normalmente usada para determinar o número ideal de clusters em um algoritmo de clusterização, como o K-means. Nela é indicada a utilização de número de *clusters* no ponto onde está o ‘cotovelo’ do gráfico.

No contexto da aplicação do algoritmo de agrupamento é fundamental avaliar o desempenho do algoritmo por meio de métricas que mensurem a qualidade da clusterização obtida. Existem diversas métricas de desempenho disponíveis para essa finalidade. Nesta pesquisa, foram empregadas algumas das métricas mais comumente utilizadas, conforme indicadas no Quadro 5.

Quadro 5: Métricas de desempenho de Clusterização

Clusterização - Métricas de desempenho	**Aplicação**
Coeficiente de Silhueta (Silhouette Coefficient):	Essa métrica mede a qualidade da separação entre os clusters. Ela varia de -1 a 1, onde valores mais próximos de 1 indicam uma separação clara e valores próximos de -1 indicam que as amostras foram atribuídas ao cluster errado (MACÊDO; SANTOS; MACIEL, 2020).
Índice de Rand ajustado (Adjusted Rand Index - ARI)	Essa métrica compara a similaridade entre os agrupamentos obtidos e os rótulos verdadeiros dos dados. O valor do índice varia de -1 a 1, onde valores mais próximos de 1 indicam uma concordância perfeita entre os agrupamentos e os rótulos verdadeiros (ULTSCH; LÖTSCH, 2022).
Índice de Pureza (Purity Score)	Essa métrica mede a pureza dos clusters, ou seja, a proporção de amostras corretamente atribuídas ao cluster majoritário. O valor do índice varia de 0 a 1, onde valores mais próximos de 1 indicam uma atribuição correta das amostras aos clusters (FUENTEALBA; LÓPEZ; PONCE, 2021).
Inércia	A inércia é a soma das distâncias quadráticas médias de cada amostra para o *centróide* do seu cluster. É uma medida interna de qualidade da clusterização, onde valores menores indicam uma clusterização mais compacta (OBERTO, 2020).

Fonte: Autora (2023)

O valor do Silhouette Score obtido foi de 0,6209221078493415. O *Silhouette Score* é uma medida de qualidade de clusterização que varia de -1 a 1, onde valores mais próximos de 1 indicam uma clusterização mais densa e bem separada, enquanto

valores próximos de -1 indicam uma clusterização inadequada. O valor obtido de 0,62 (Figura 23) sugere que a clusterização realizada apresenta uma boa separação entre os clusters (MACÊDO; SANTOS; MACIEL, 2020).

Figura 23: Silhouette Score

```
# Silhouette Score
labels = modelo_v1.labels_
silhouette_score(pca, labels, metric = 'euclidean')
```

```
0.6209221078493415
```

Fonte: Autora (2023)

A próxima métrica analisada foi Índice de Rand ajustado (Adjusted Rand Index - ARI). Se o resultado for 1.0, isso significa que os rótulos atribuídos pelo seu modelo de clusterização são perfeitamente consistentes com os rótulos verdadeiros dos clusters. Um valor de ARI igual a 1.0 indica uma correspondência perfeita entre os rótulos preditos e os rótulos verdadeiros, o que é considerado um resultado muito bom (ULTSCH; LÖTSCH, 2022).

Na Figura 24 é exposto o valor obtido.

Figura 24: Índice de Rand ajustado (Adjusted Rand Index - ARI)

```
ari = adjusted_rand_score(labels_true, labels_pred)
print("Adjusted Rand Index:", ari)
```

```
Adjusted Rand Index: 1.0
```

Fonte: Autora (2023)

O *Purity Score* é uma métrica que mede a qualidade da clusterização em relação às classes verdadeiras dos dados. O valor do *Purity Score* varia de 0 a 1, onde um valor mais próximo de 1 indica uma clusterização com alta pureza, ou seja, os *clusters* estão bem alinhados com as classes verdadeiras dos dados. No caso, um *Purity Score* de 0.875 (Figura 25) é considerado um resultado bom, uma vez que

significa que aproximadamente 87,5% dos pontos foram atribuídos ao *cluster* correto, de acordo com as classes verdadeiras dos dados (FUENTEALBA; LÓPEZ; PONCE, 2021).

Figura 25: Índice de Pureza (Purity Score)

```
purity = purity_score(y_true, y_pred)
print("Purity Score:", purity)
```

```
Purity Score: 0.875
```

Fonte: Autora (2023)

A métrica de inércia é calculada no algoritmo *K-means* e representa a soma das distâncias quadradas entre cada ponto de dado e o *centróide* do cluster ao qual ele pertence. Um valor baixo de inércia indica que os pontos dentro de cada *cluster* estão próximos uns dos outros, ou seja, eles têm uma alta similaridade entre si. Neste caso, o valor de inércia foi 9,333 (Figura 26), o que é considerado relativamente baixo (OBERTO, 2020). Isso sugere que os clusters formados pelo algoritmo estão bem compactos, ou seja, os pontos dentro de cada *cluster* estão próximos uns dos outros. Essa compactação dos clusters indica que os pontos compartilham características semelhantes e estão mais distantes de outros clusters, o que é um bom sinal de uma clusterização eficiente.

Figura 26: Inércia

```
# Criação do objeto KMeans com o número de clusters desejado
kmeans = KMeans(n_clusters=2)

# Ajuste do modelo aos dados
kmeans.fit(X)

# Obtenção da inércia
inertia = kmeans.inertia_

print("Inertia:", inertia)
```

```
Inertia: 9.333333333333334
```

Fonte: Autora (2023)

Com base nos resultados apresentados é possível inferir que a clusterização obteve um desempenho muito bom. No Quadro 6 são expostos os resultados sumarizados das métricas do algoritmo K-means.

Quadro 6: Análise dos resultados do algoritmo - K-means

Métrica	**Resultado**
Silhouette Score	0,6209221078493415
Índice de Rand ajustado (*Adjusted Rand Index - ARI*)	Adjusted Rand Index: 1,0
Índice de Pureza (*Purity Score*)	Purity Score: 0,875
Inércia	Inertia: 9,333333333333334

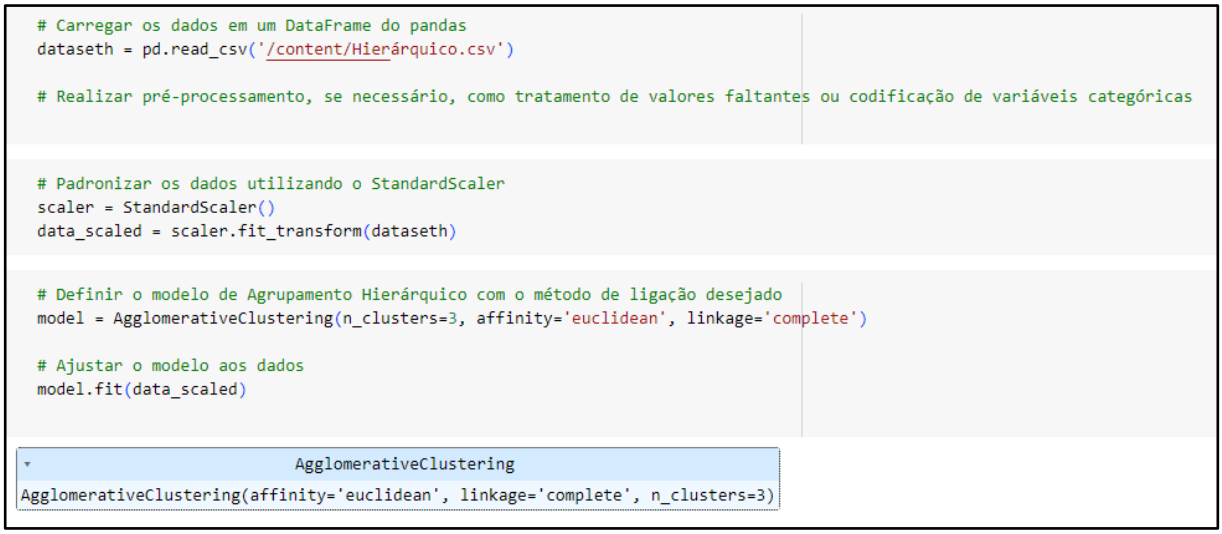
Fonte: Autora (2023)

Em resumo, os resultados obtidos indicam que a clusterização teve um desempenho ótimo. Esses resultados confirmam que a clusterização foi bem-sucedida em identificar padrões e agrupar os dados com base em suas similaridades. Portanto, conclui-se que a clusterização obteve um desempenho ótimo, proporcionando *clusters* compactos, de alta pureza e concordância perfeita com as classes verdadeiras dos dados.

3.5.2 Agrupamento Hierárquico

Neste capítulo são apresentados os resultados da aplicação do algoritmo *Hierarchical Clustering*. O principal objetivo do agrupamento hierárquico é agrupar os dados de forma hierárquica, ou seja, criar uma hierarquia de *clusters* que podem ser visualizados em forma de dendrograma. Isso permite que os dados sejam organizados em diferentes níveis de granularidade, desde clusters grandes e mais gerais, até clusters menores e mais específicos. O algoritmo de agrupamento hierárquico aplicado é exposto na Figura 27.

Figura 27: Agrupamento Hierárquico



Fonte: Autora (2023)

No Quadro 7 são expostos os atributos envolvidos nas cinco clusterizações realizadas nesta pesquisa.

Quadro 7: Clusterização realizada pelo algoritmo - Hierarchical clustering

Clusterização	Atributos
1	Gênero x Idade
2	Distribuição Ensino Médio (EEMN x EEMT x EEME)
3	Distribuição Médias (MFG x MFT x MF)
4	Distribuição Médias (Aprovado x Reprovado x Trancou o Curso)
5	CC x SI x TADS x TGTI x TRC x TSEG x TSIN

Fonte: Autora (2023)

3.5.3 Agglomerative Clustering (Agrupamento Aglomerativo)

Neste tópico são apresentados os resultados da aplicação do algoritmo *Agglomerative Clustering*. O *Agglomerative Clustering* (Agrupamento Aglomerativo) é um algoritmo de aprendizado não supervisionado utilizado para agrupar objetos em *clusters*. Ele começa considerando cada objeto como um *cluster* separado e, em seguida, mescla iterativamente os *clusters* mais próximos uns aos outros com base em uma métrica de distância, formando um dendrograma.

O objetivo do *Agglomerative Clustering* cria uma hierarquia de agrupamentos, sendo sua premissa que objetos que estão mais próximos uns dos outros são mais similares do que objetos que estão mais distantes. Na Figura 28 é apresentada o algoritmo elaborado para esta finalidade.

Figura 28: Agglomerative Clustering (Agrupamento Aglomerativo)

```
from sklearn.cluster import AgglomerativeClustering
from sklearn.preprocessing import StandardScaler

# Pré-processamento dos dados, se necessário
# X = ...

# Padronização dos dados
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# Criando o modelo de clustering hierárquico
# Neste exemplo, usaremos o método de ligação completa (complete linkage)
# Você pode escolher outros métodos, como ligação simples (single linkage) ou ligação média (average linkage)
model = AgglomerativeClustering(n_clusters=3, linkage='complete')

# Ajuste do modelo aos dados
model.fit(X_scaled)

# Obtenção das etiquetas dos clusters atribuídos a cada ponto de dado
labels = model.labels_

# Imprimir as etiquetas atribuídas a cada ponto de dado
print(labels)
```

Fonte: Autora (2023)

Após importação e ajuste do algoritmo *Agglomerative Clustering* faz-se necessário realizar a redução de dimensionalidade. Para tanto, foi aplicada a análise PCA *(Principal Component Analysis*), conforme apresentado na Figura 29.

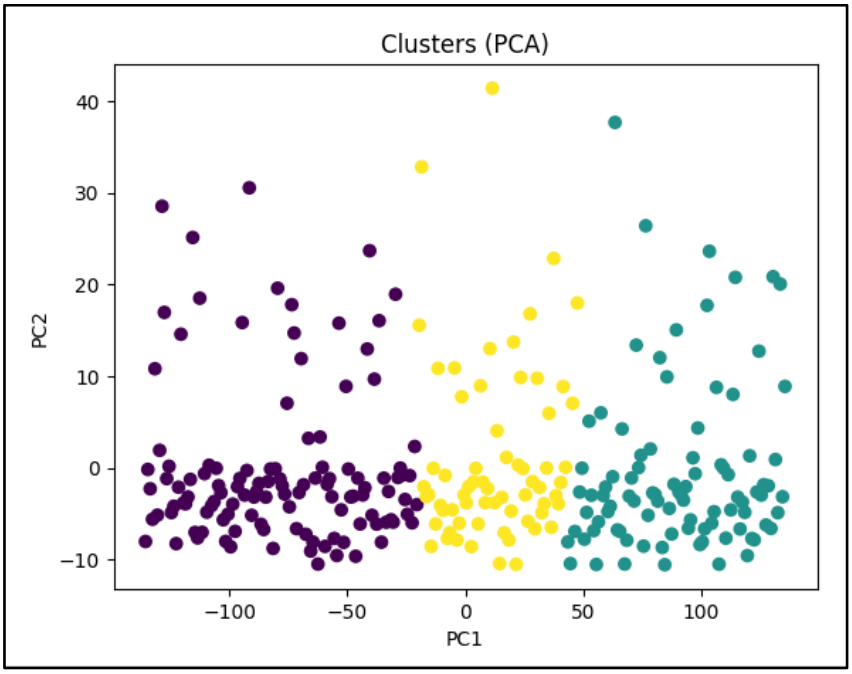
Figura 29: Agglomerative Clustering (Agrupamento Aglomerativo)

```
from sklearn.decomposition import PCA

# Aplicar PCA para reduzir a dimensionalidade para 2 componentes
pca = PCA(n_components=2)
data_pca = pca.fit_transform(dataset2.drop('Cluster', axis=1))

# Plotar um gráfico de dispersão dos clusters após redução de dimensionalidade
plt.scatter(data_pca[:, 0], data_pca[:, 1], c=dataset2['Cluster'])
plt.xlabel('PC1')
plt.ylabel('PC2')
plt.title('Clusters (PCA)')
plt.show()
```

Fonte: Autora (2023)

O PCA é uma técnica que visa reduzir o número de variáveis em um conjunto de dados, mantendo a maior quantidade possível de informações relevantes. O PCA é amplamente utilizado em análise de dados e aprendizado de máquina para lidar com conjuntos de dados de alta dimensionalidade. Ele possui várias aplicações e benefícios, incluindo: Redução da complexidade dos dados; Identificação de padrões e estrutura nos dados; Eliminação de variáveis redundantes; Visualização de dados em espaços de menor dimensão; Pré-processamento de dados para algoritmos de aprendizado de máquina: A redução de dimensionalidade pode melhorar a eficiência computacional e evitar problemas de overfitting em modelos que sofrem com a maldição da dimensionalidade (JAFARZADEGAN, *et al.*, 2019). Na Figura 30 é apresentado um exemplo de gráfico para a visualização dos *clusters* produzido pelo PCA.

Figura 30: Visualização dos clusters obtidos com PCA



Fonte: Autora (2023)

A Figura acima apresenta a visualização dos clusters após a aplicação do PCA. É possível identificar que os dados foram agrupados em 3 *clusters* principais, representados pelas cores roxa, amarelo e verde. As aplicações indicadas permitem, para os fins desta pesquisa, a criação de indicadores e a realização de comparações mais precisas e eficazes entre diferentes grupos de dados dos alunos, facilitando assim a identificação de padrões, tendências e áreas de melhoria na formação

superior. Além disso, os dados transformados podem ser utilizados na geração de gráficos, relatórios e outras representações visuais que auxiliam na compreensão e na tomada de decisões informadas pela gestão acadêmica.

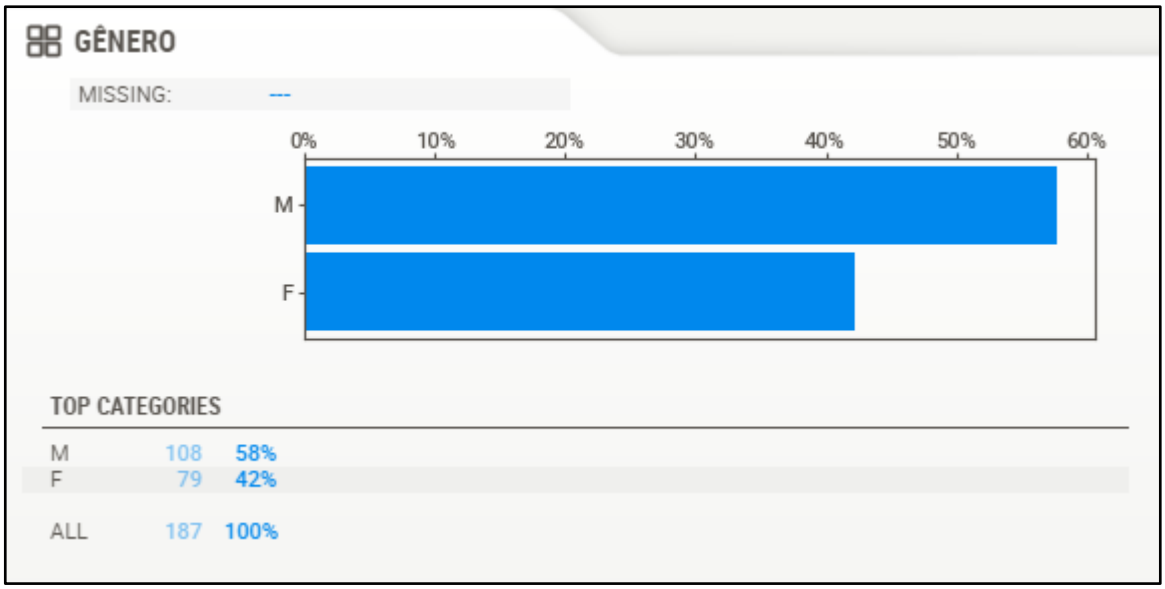
4. APRESENTAÇÃO E ANÁLISE DOS RESULTADOS DA SOLUÇÃO AUTOMATIZADA

Neste capítulo são apresentados os resultados obtidos com a aplicação da solução automatizada para ambos os cursos analisados (Governança em TI e Data Science), a partir das análises feitas na ferramenta *Google Colab* e pelo software de análise de dados acadêmicos *Tableau*. Nas Figuras 20, 21, 22 e 23, expostas nos tópicos a seguir, é apresentado um panorama da base de dados com suas porcentagens, considerando-se cada categoria de dado analisada. A partir dos resultados da aplicação dos algoritmos de predição indicados prossegue-se para a etapa de geração de relatórios e gráficos. As informações desses dashboards auxiliarão o gestor a compreender o perfil dos alunos e assim identificar o prognóstico de perfil ideal para oferta de cursos específicos, de maneira direcionada e assertiva aos futuros candidatos, conforme a sequência de quatro fases indicada no capítulo anterior.

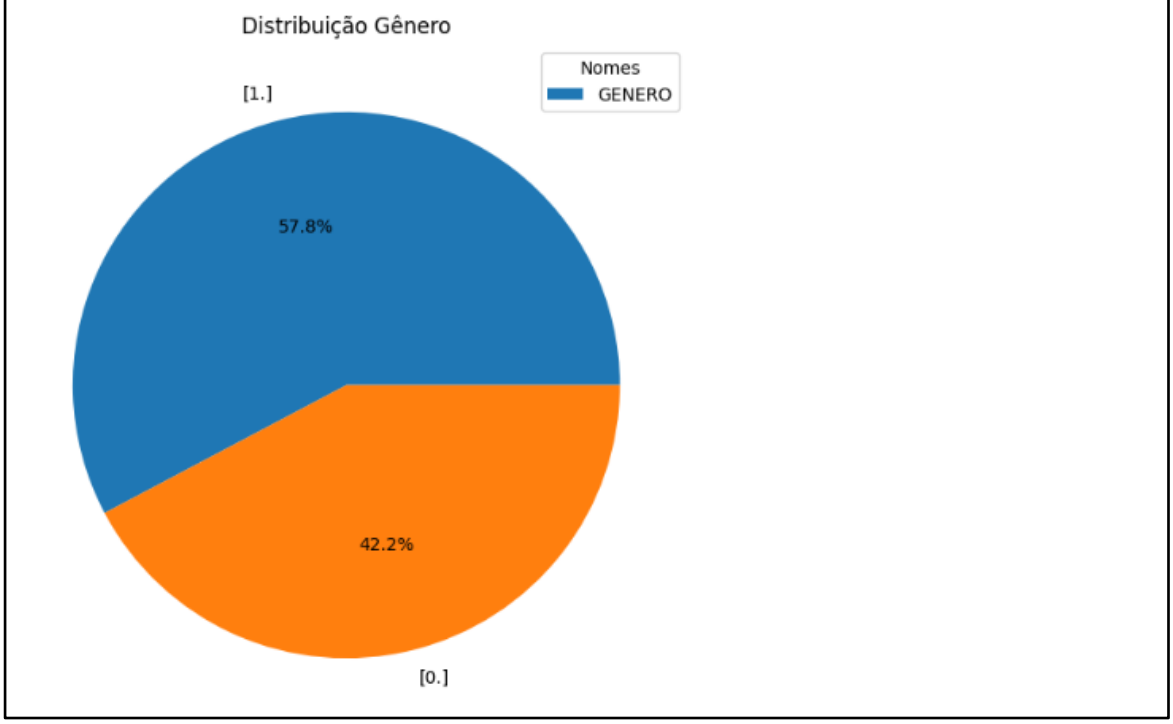
4.1 Resultados do Curso Governança em TI

Nas figuras a seguir são apresentadas as análises obtidas da ferramenta *Google Colab* para o curso de Governança em TI. De início, a Figura 20 mostra a proporção entre alunos, conforme o atributo ‘gênero’.

Figura 31: Distribuição dos alunos no atributo ‘gênero’



```
plt.figure(figsize=(7, 7))  # Define as dimensões da figura (largura x altura)
plt.pie(dataseth[['GENERO']].value_counts(), labels=dataseth[['GENERO']].drop_duplicates().values, autopct='%1.1f%%')
plt.title('Distribuição Gênero')
plt.legend(['GENERO'], title='Nomes')
plt.show()
```

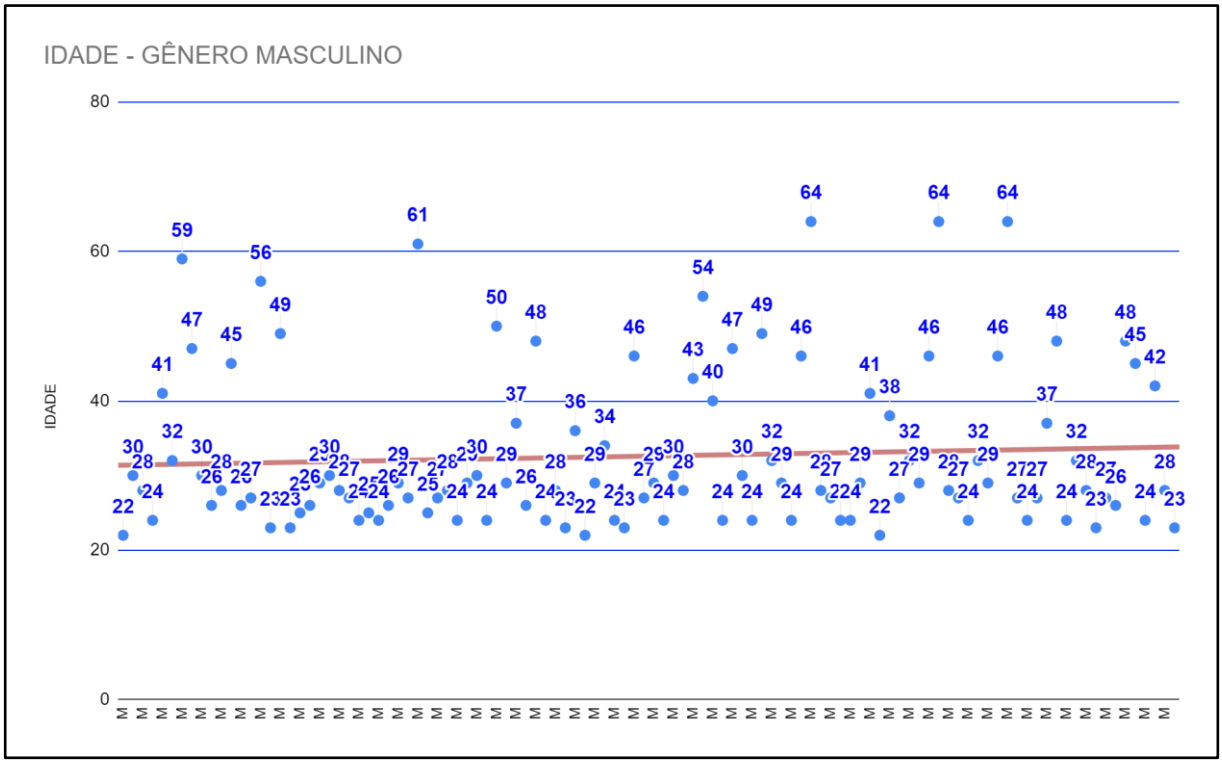


Fonte: Autora (2023)

A Figura 31 exibe a distribuição proporcional de alunos por gênero, destacando uma prevalência masculina. Dos 187 alunos no total, observa-se que 108 são do sexo masculino, representando aproximadamente 57,8% da população, enquanto 79 são do sexo feminino, correspondendo a cerca de 42,2%.

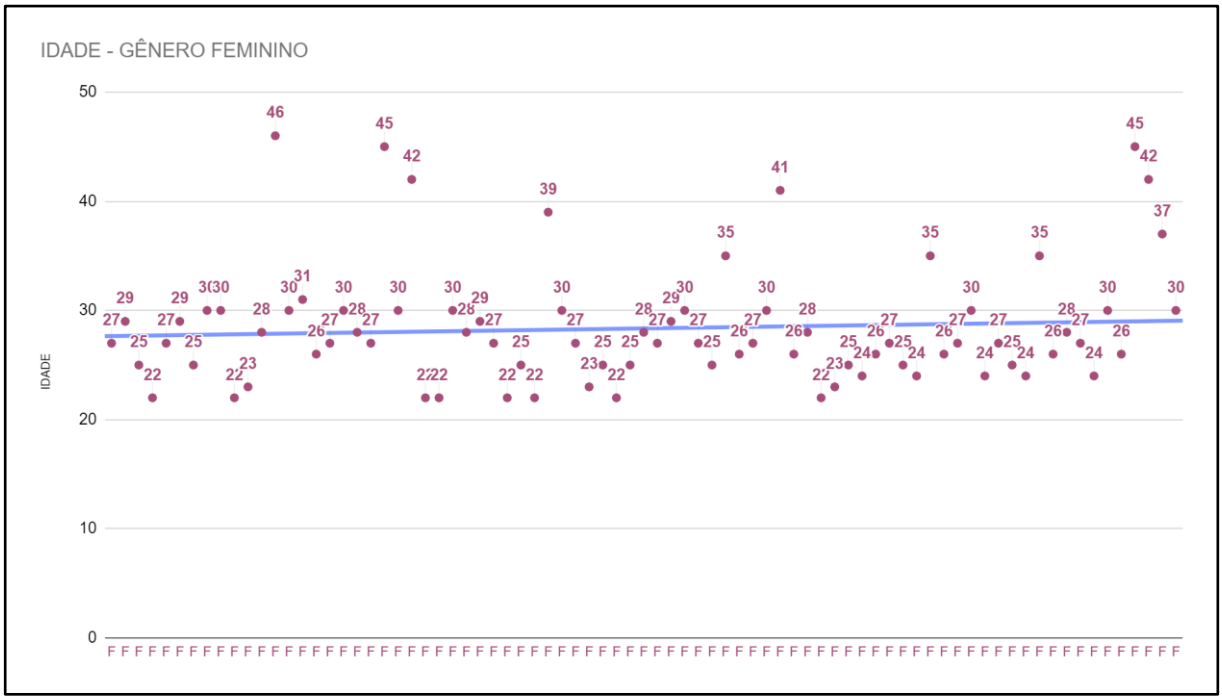
Nas Figuras 32 e 33 são apresentadas visualizações das idades dos alunos, separados por gênero. Inicialmente são exibidas as idades dos alunos do gênero masculino, seguidas pelas idades dos alunos do gênero feminino. Essa representação gráfica permite observar de forma clara e comparativa a distribuição das idades entre os dois grupos.

Figura 32: Distribuição dos alunos no atributo gênero - Masculino



Fonte: Autora (2023)

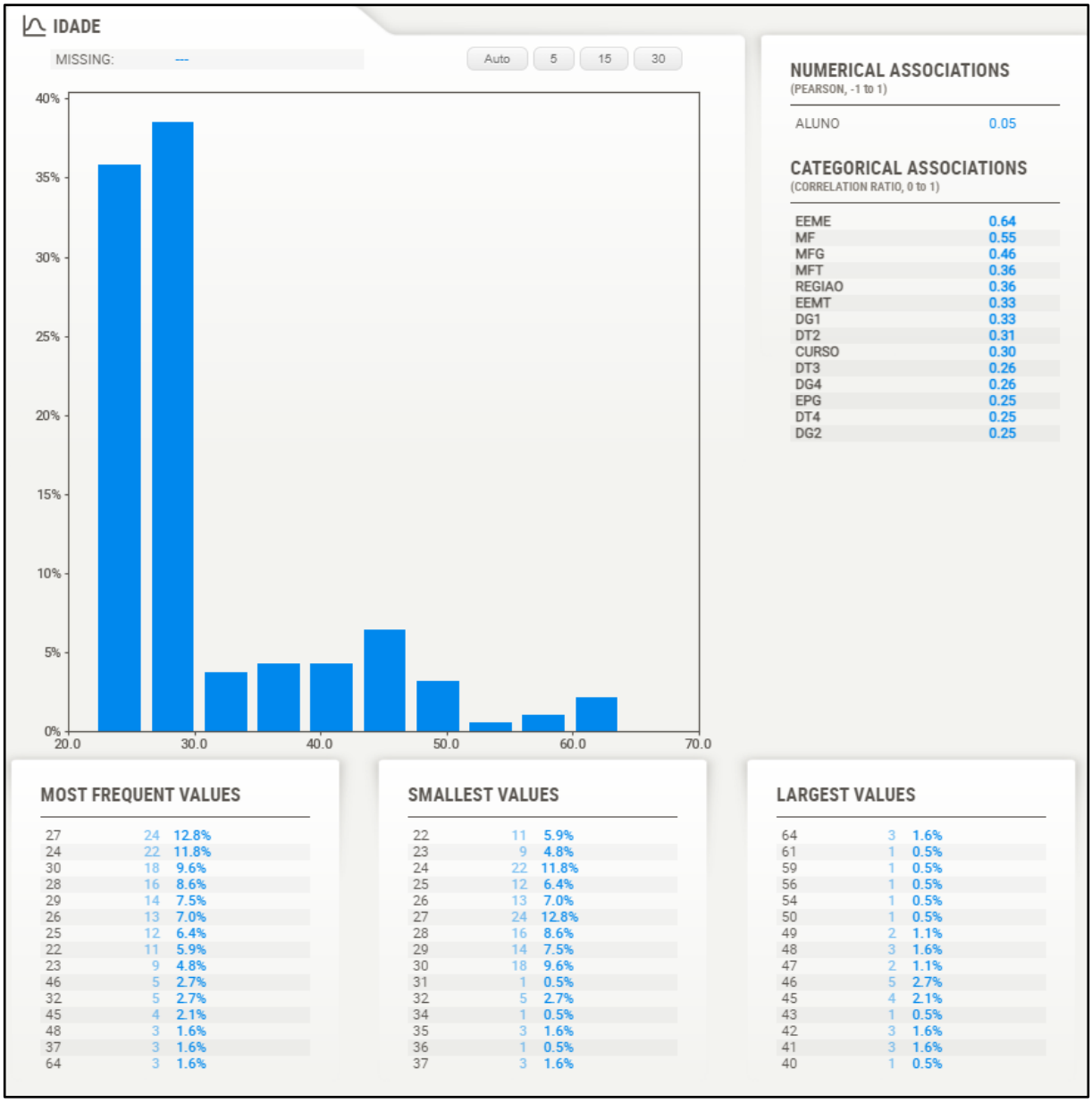
Figura 33: Distribuição dos alunos no atributo gênero - Feminino



Fonte: Autora (2023)

As Figuras 32 e 33 apresentam a distribuição de idades de uma população de alunos, separados por gênero masculino e feminino. No grupo masculino, observa-se que as idades variam de 22 a 64 anos, com uma média aritmética de 32 anos. Já no grupo feminino, as idades variam entre 22 e 46 anos, com uma média aritmética de 28 anos. A média aritmética é calculada pela soma das idades de todos os indivíduos do grupo, dividida pelo número total de indivíduos. Essa medida estatística fornece uma estimativa do valor central das idades em cada grupo, permitindo uma compreensão resumida da distribuição etária dos alunos. A Figura 34 exibe a distribuição dos alunos por idade.

Figura 34: Distribuição dos alunos no atributo - Idade

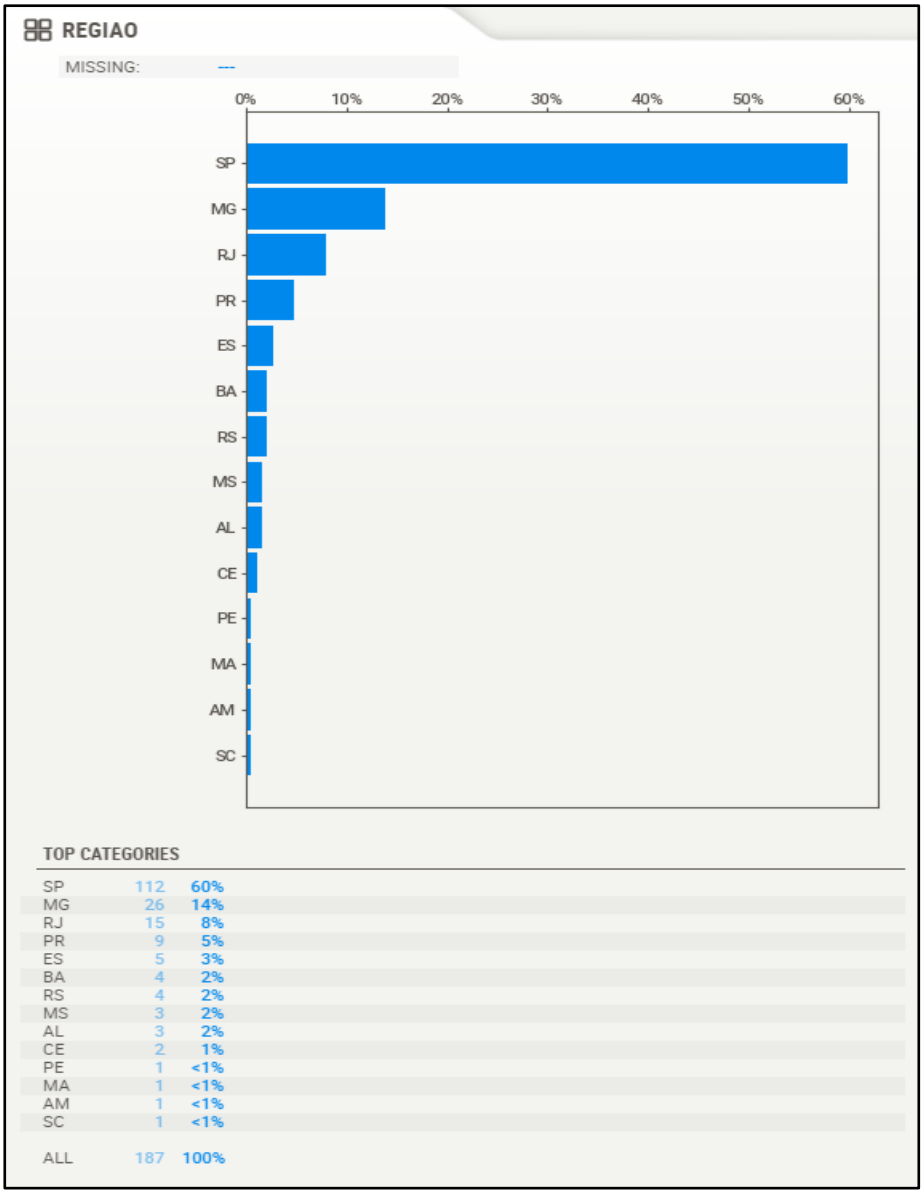


Fonte: Autora (2023)

Os dados revelam que a faixa etária mais comum entre os alunos é de 24 a 30 anos. Especificamente, 12,8% dos alunos têm 27 anos; 11,8% têm 24 anos; 9,6% têm 30 anos; 8,6% têm 28 anos e 7,5% têm 29 anos. Esses resultados indicam uma tendência de procura mais acentuada por jovens recém-formados na graduação, buscando especialização em sua área de estudo. Já alunos acima de 50 anos não são um público muito frequente para este curso de pós-graduação, pois representam cerca de 1% a 2% da população do curso.

Na Figura 35 são apresentados os resultados do atributo ‘região’ (estado de domicílio no Brasil).

Figura 35: Distribuição dos alunos no atributo ‘região’

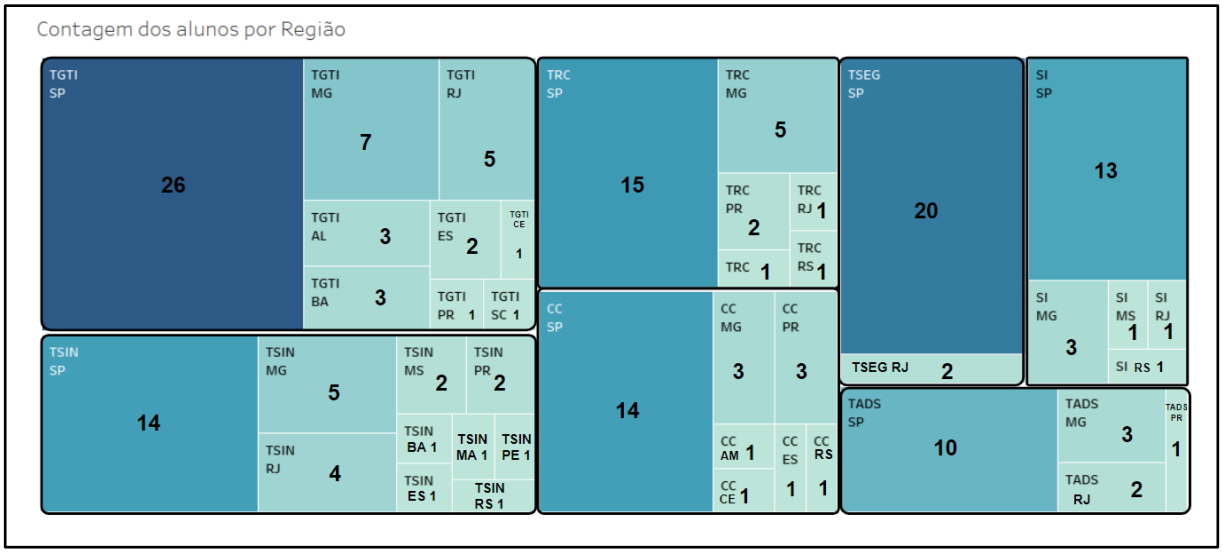


Fonte: Autora (2023)

Observa-se uma predominância significativa de alunos provenientes da região São Paulo, com 112 alunos residentes nesse estado, representando aproximadamente 60% do total de matrículas na pós-graduação em Governança de TI. Em segundo lugar está o estado de Minas Gerais com 26 alunos, correspondendo a 14% e em terceiro lugar o estado do Rio de Janeiro com 15 alunos, representando 8%, em quarto lugar o estado de Paraná com 9 alunos, representando 5% e em quinto

lugar o estado de Espírito Santo com 5 alunos, aproximadamente 3% do total analisado.

Figura 36: Distribuição dos alunos no atributo ‘região’ e por curso

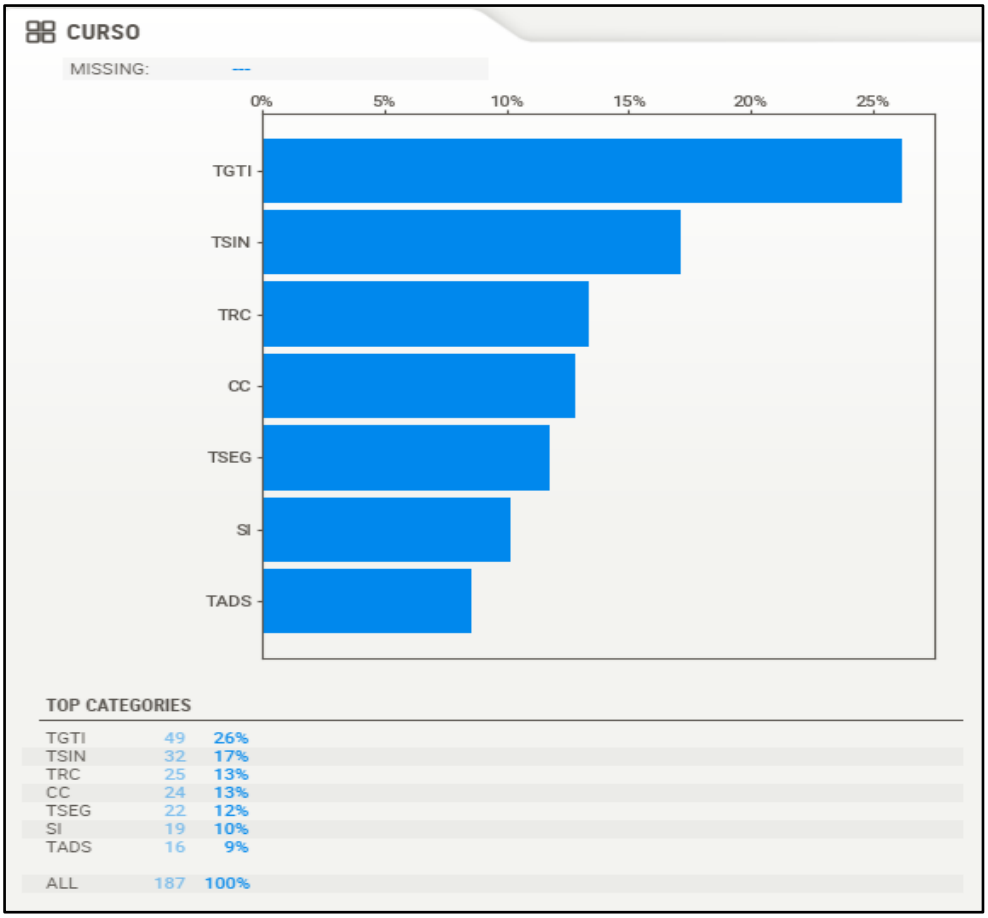


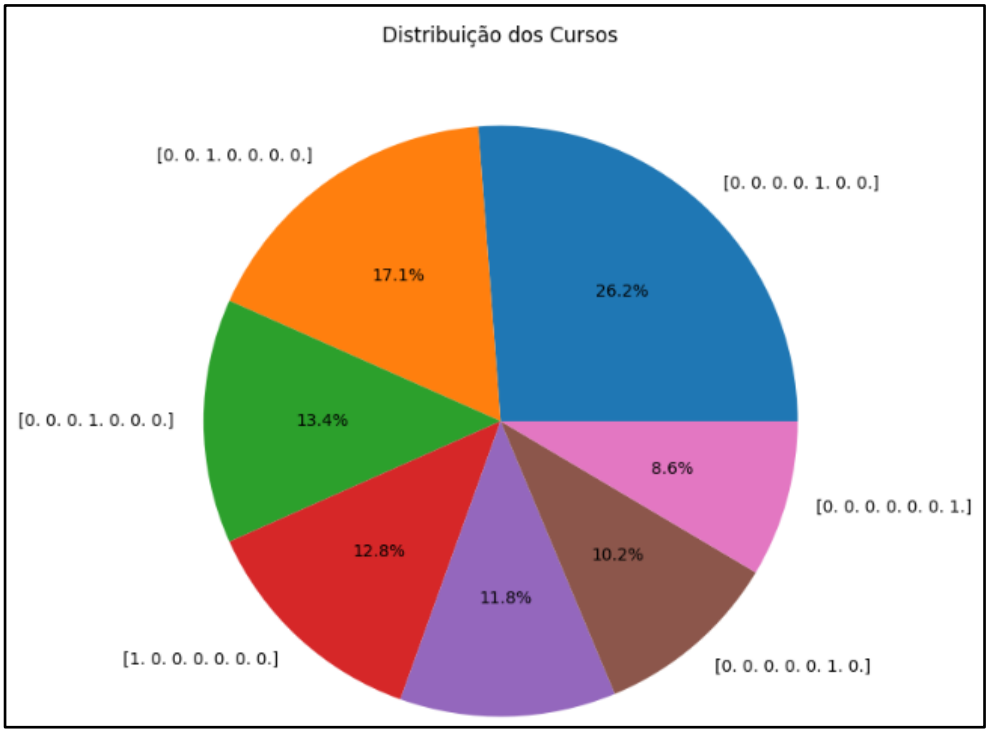
Fonte: Autora (2023)

Na Figura 36 foi apresentado a distribuição de alunos por curso e Região do país, onde fica evidente novamente a soberania do estado de São Paulo, diante aos demais estados brasileiros neste atributo ‘Região’. Neste atributo ainda é possível notar que o curso de TGTI, em vista ao estado de São Paulo possui 26 alunos dos 112 analisados desta região, o que equivale a 23,2%, tendo o segundo lugar o curso de TSEG com 20 alunos representando 17,9% e por fim, TRC com 15 alunos obtendo a 13,3% dos alunos analisados nesta população do estado de São Paulo. Analisando este parâmetro, o curso de TADS mostra-se o menor entre os demais com apenas 10 alunos residentes em São Paulo, representando aproximadamente 9%.

Na Figura 37 são apresentados os resultados da correlação entre os atributos ‘idade’ e ‘curso de graduação’

Figura 37: Análise do atributo ‘curso de graduação’



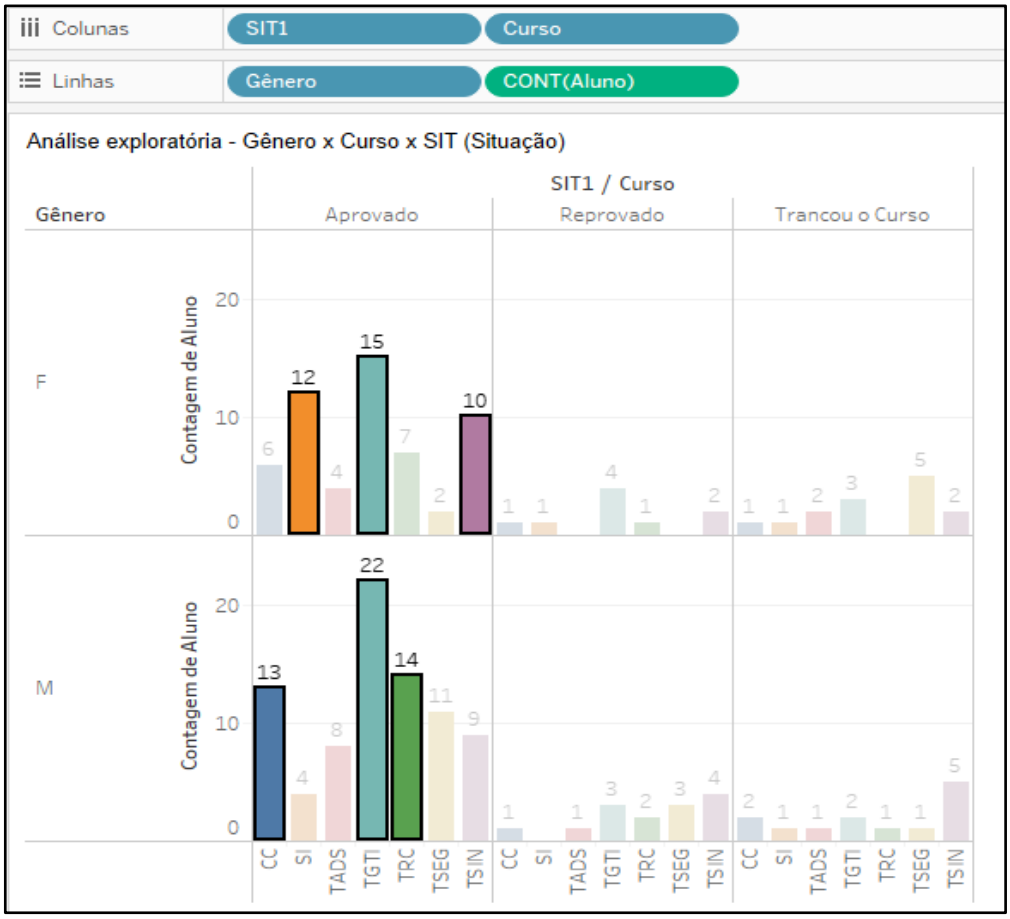


Fonte: Autora (2023)

Na Figura 37 é possível visualizar as percentagens referentes ao atributo ‘curso de graduação’, onde se acentua a quantidade de alunos formados no curso de graduação de TGTI (Tecnologia em Gestão de Tecnologia da Informação), com 49 dos 187 alunos, o que corresponde a 26%. Verifica-se ainda que a menor quantidade de alunos tem como formação a graduação de TADS (Tecnologia em Análise e Desenvolvimento de Sistemas), com apenas 16 dos 187 alunos, o que corresponde a 9% do total analisado.

Com a ferramenta do *Google Colab* foram realizadas análises comparativas entre os resultados dos atributos de dados expostos anteriormente, visando extrair padrões de perfil de aluno do curso de pós-graduação em relação à sua performance no curso. Na Figura 26, utilizando-se o software *Tableau,* são analisados de forma cruzada os atributos ‘gênero’, ‘curso de graduação’ e ‘situação’ do aluno.

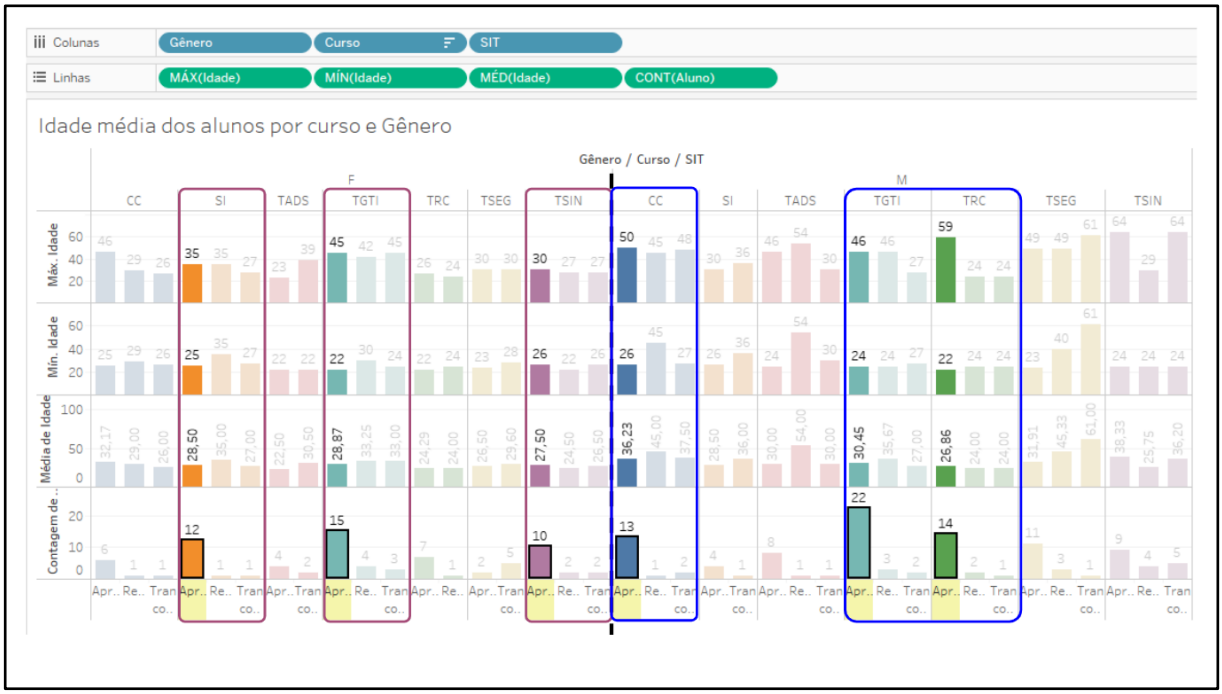
Figura 38: Correlação entre os atributos gênero, curso de graduação e situação



Fonte: Autora (2023)

É possível perceber a maior sucesso ao curso de pós-graduação de alunos egressos do curso de graduação em TGTI (Tecnologia em Gestão de Tecnologia da Informação) em ambos os gêneros, com 22 alunos do gênero masculino e 15 alunos do gênero feminino. Tal resultado demonstra maior procura por curso de pós-graduação (Governança em TI) na mesma linha da formação do aluno no curso de graduação concluído (TGTI). Já os alunos egressos dos cursos de graduação em SI (Sistemas de Informação) e TSIN (Tecnologia em Sistemas para Internet) são predominantemente do gênero feminino, com 12 e 10 alunas aprovados respectivamente; enquanto os cursos TRC (Tecnologia em Redes de Computadores) e CC (Ciência da Computação) têm predominância de alunos do gênero masculino, com 14 e 13 alunos respectivamente. Na Figura 39 é apresentada esta relação adicionando o atributo ‘idade’.

Figura 39: Correlação entre os atributos gênero, curso de graduação e situação com Idade



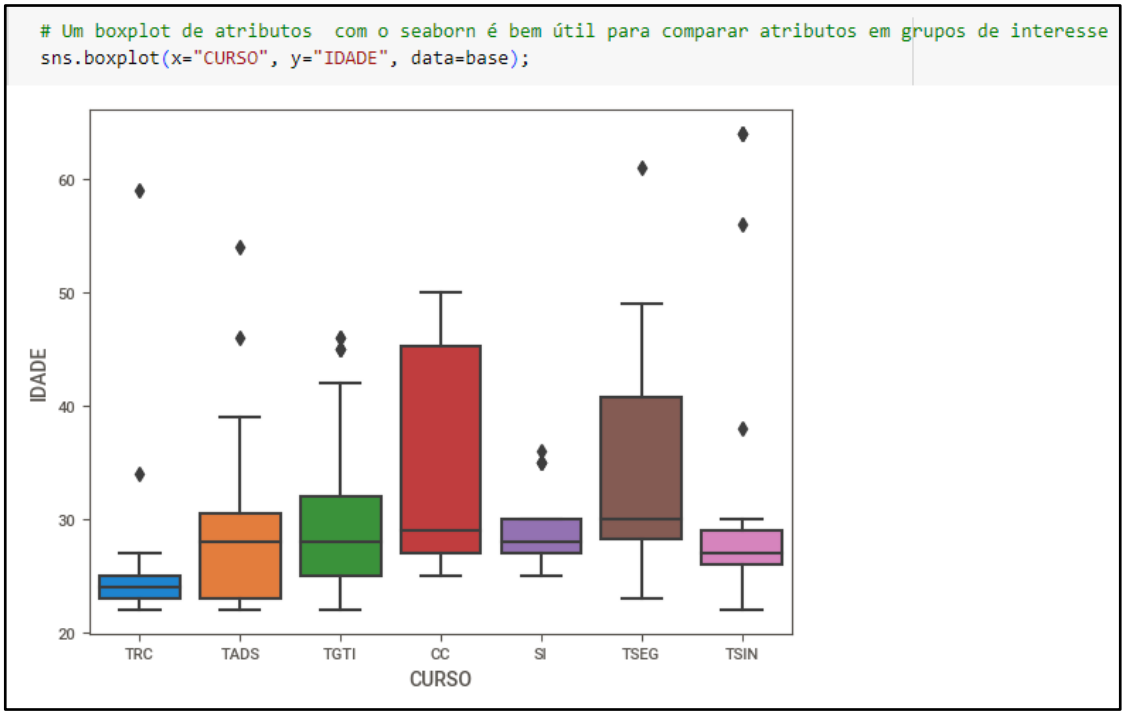
Fonte: Autora (2023)

A Figura 39 apresenta a distribuição de alunos considerando-se os atributos gênero, curso de graduação, situação e idade. Verificou-se que 11,8% dos alunos do gênero masculino obtiveram o maior sucesso no curso de TGTI, totalizando 22 alunos aprovados, com idades entre 22 e 46 anos. De maneira semelhante, as alunas do gênero feminino representaram 8% das aprovações no mesmo curso, com um total

de 15 alunas aprovadas, com idades entre 22 e 45 anos, que nos indica que este curso possui um público bem variável considerando este intervalo de idades. Com olhar ainda no público feminino curso de SI representa 6,4% das aprovações, com um total de 12 alunas aprovadas, com idades entre 25 e 35 anos, um público um pouco mais jovem em comparação ao curso anterior, em TSIN temos 5,3% das alunas aprovadas, com 10 alunas em idade entre 26 e 30 anos. Tal resultado expõe um intervalo de idades menor em relação ao curso com melhor desempenho.

Do lado masculino há, além do curso que obteve maior destaque (TGTI), o curso de TRC com 7,5%, somando 14 alunos, de idades entre 22 e 59 anos. O que leva a entender que este curso possui egressos com idades mais elevadas em comparação aos outros cursos. Aparece também neste público masculino, o curso de CC com idades entre 26 e 50 anos, que mostra ser um perfil bem variável neste aspecto, totalizando 13 aprovações, que equivale a 7% deste grupo estudado. A Figura 40 apresenta, de forma mais detalhada, o atributo idade, permitindo uma visualização mais clara dos intervalos de idades dos cursos.

Figura 40: Correlação entre os atributos gênero, curso de graduação e situação com Idade



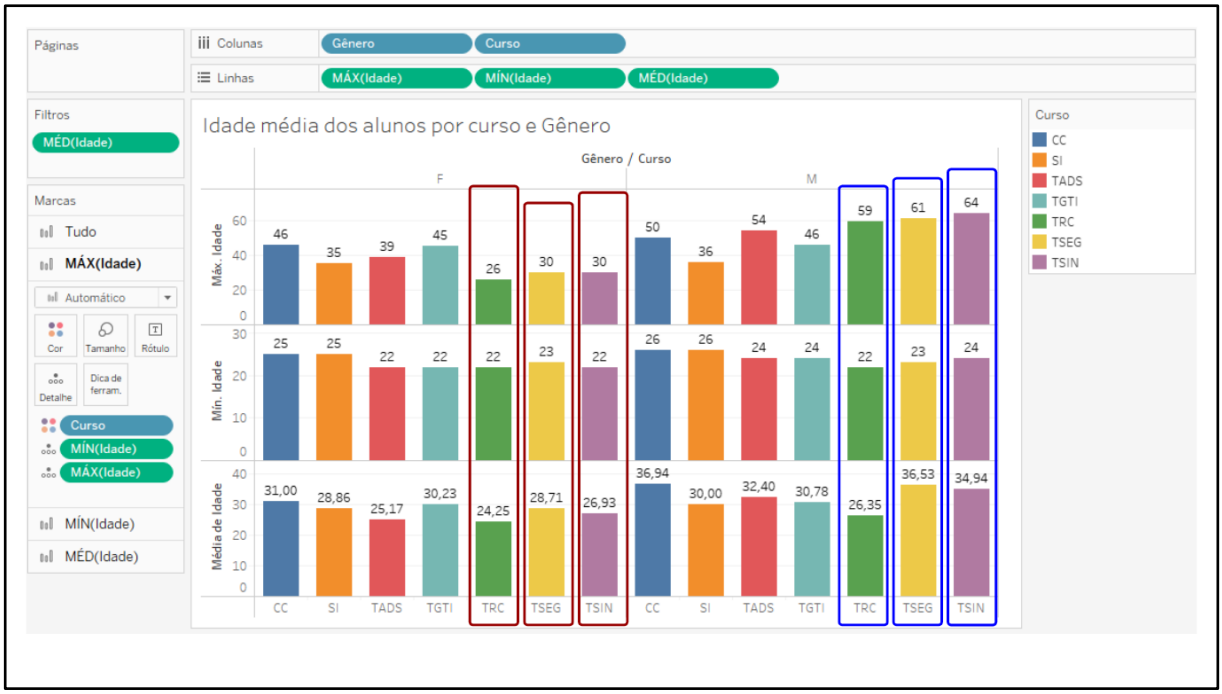
Fonte: Autora (2023)

A comparação dos atributos ‘idade’ e ‘curso de graduação’ indica que os alunos mais jovens concentram-se nos cursos de TRC (Tecnologia em Redes de

Computadores), onde 20 alunos entre os 25 egressos deste curso, representando a faixa etária de maior expressão entre 22 a 25 anos, o que equivale a 80% dos alunos matriculados vindos deste curso. No curso de SI (Sistemas de Informação), de 19 alunos matriculados 15 possuem entre 25 e 30 anos, o que remete a 78,95% dos alunos. Já no curso TSIN (Tecnologia em Sistemas para Internet) os alunos apresentam a maior concentração na faixa entre 22 a 28 anos, equivalentes a 65,62% de alunos egressos deste curso. O curso de CC (Ciência da Computação), contém 11 de 24 alunos na faixa entre 32 e 50 anos. Por fim, o curso de TSEG (Tecnologia em Segurança da Informação) possui 13 de 22 alunos com idades entre 30 e 61 anos, o que representa 59,09% da população vinda deste curso.

Na Figura 41 são apresentados os resultados das análises efetuadas, considerando-se os atributos idade e curso de graduação, porém, com relação aos gêneros feminino e masculino dos alunos analisados.

Figura 41: Correlação entre os atributos ‘curso de graduação’, ‘idades’ e ‘gênero’



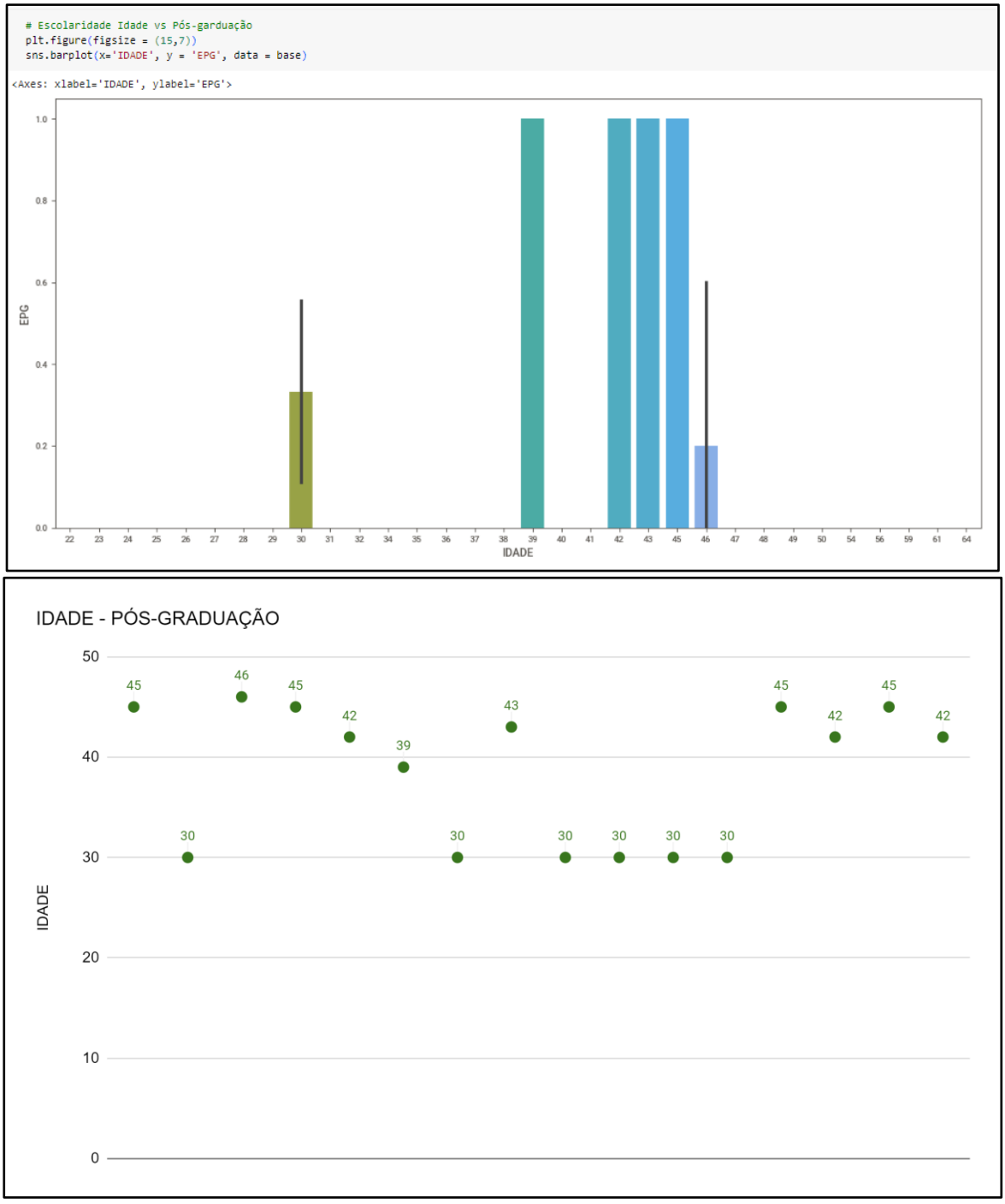
Fonte: Autora (2023)

Realizando-se uma análise da distribuição etária da população geral de alunos, separada por curso e gênero, observa-se que os estudantes do gênero masculino provenientes dos cursos de TSIN apresentam alunos com idades mais elevadas em relação aos demais cursos, gerando um perfil de alunos com até 64 anos. Em

sequência há os alunos do curso de TSEG, com até 61 anos e, por último, do curso TRC, com alunos até 59 anos. Isso evidencia que os cursos de tecnólogo estão atraindo alunos com maior propensão a buscar uma especialização após sua conclusão.

Por outro lado, esses mesmos cursos são compostos por um perfil de alunas (gênero feminino) de idades mais baixas, com alunas de até 30 anos nos cursos de TSEG e TSIN, considerando-se que no curso de TSIN o grupo de alunas equivale a 26,2% da população e 13,1% no curso de TSEG. Já no curso de TRC o perfil de alunas até 26 anos representa 15% da população da base de dados utilizada nesta pesquisa. Tal resultado evidencia que o perfil de alunas de maior idade está localizado em cursos de bacharelado, tais como Ciências da Computação (CC), com alunas entre 25 e 46 anos, que representam 15% da população; seguido pelo curso de Tecnologia em Gestão de Tecnologia da Informação (TGTI), com 41,1% da população estudada, de alunas até 45 anos. Na Figura 42 são apresentados os atributos referentes aos alunos que já possuem uma pós-graduação.

Figura 42: Correlação entre os atributos, ‘curso de Pós-graduação’ e ‘idades’

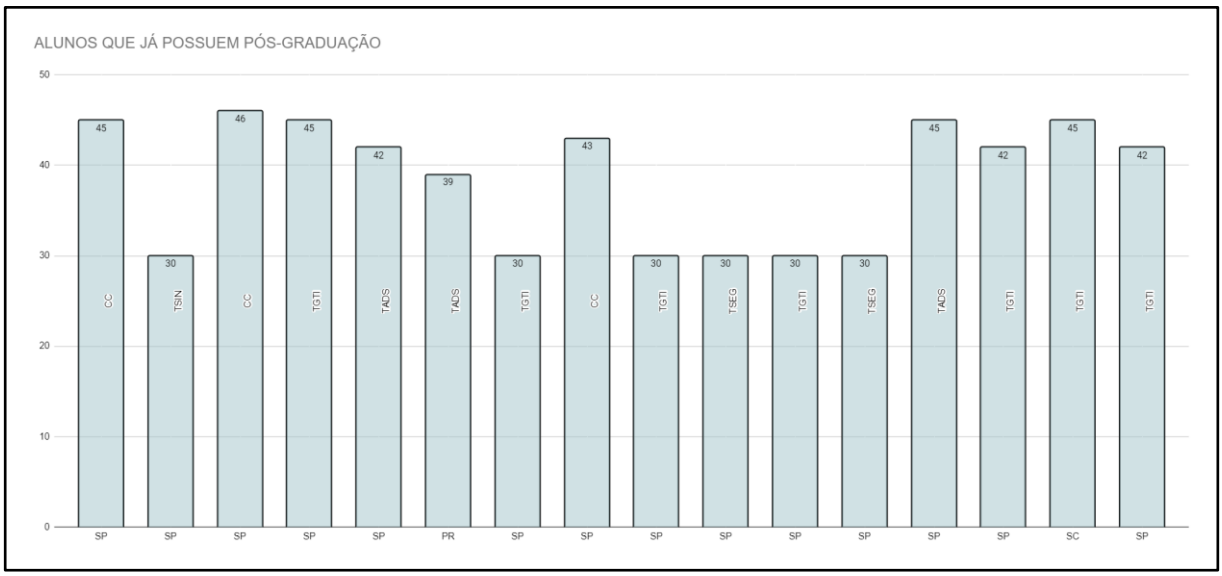


Fonte: Autora (2023)

Nota-se na Figura 42 que os resultados do grupo de alunos que estão cursando uma especialização pela segunda vez contam com 16 alunos. Deste total pode-se observar que a maioria, neste caso 6 alunos, possuem 30 anos, o que equivale a 37,5% deste grupo. Em segundo aparecem os alunos de 45 anos, 4 ao todo, o que representam 25,0% da população e em terceiro os alunos de 42 anos, que somam 3 alunos, correspondendo a 18,8% do total. Na Figura 43 visualiza-se a segregação

destes alunos a partir dos atributos região e cursos de graduação do qual são egressos.

Figura 43: Correlação entre os atributos ‘curso de Pós-graduação’, ‘idade’, ‘região’ e ‘curso de graduação’



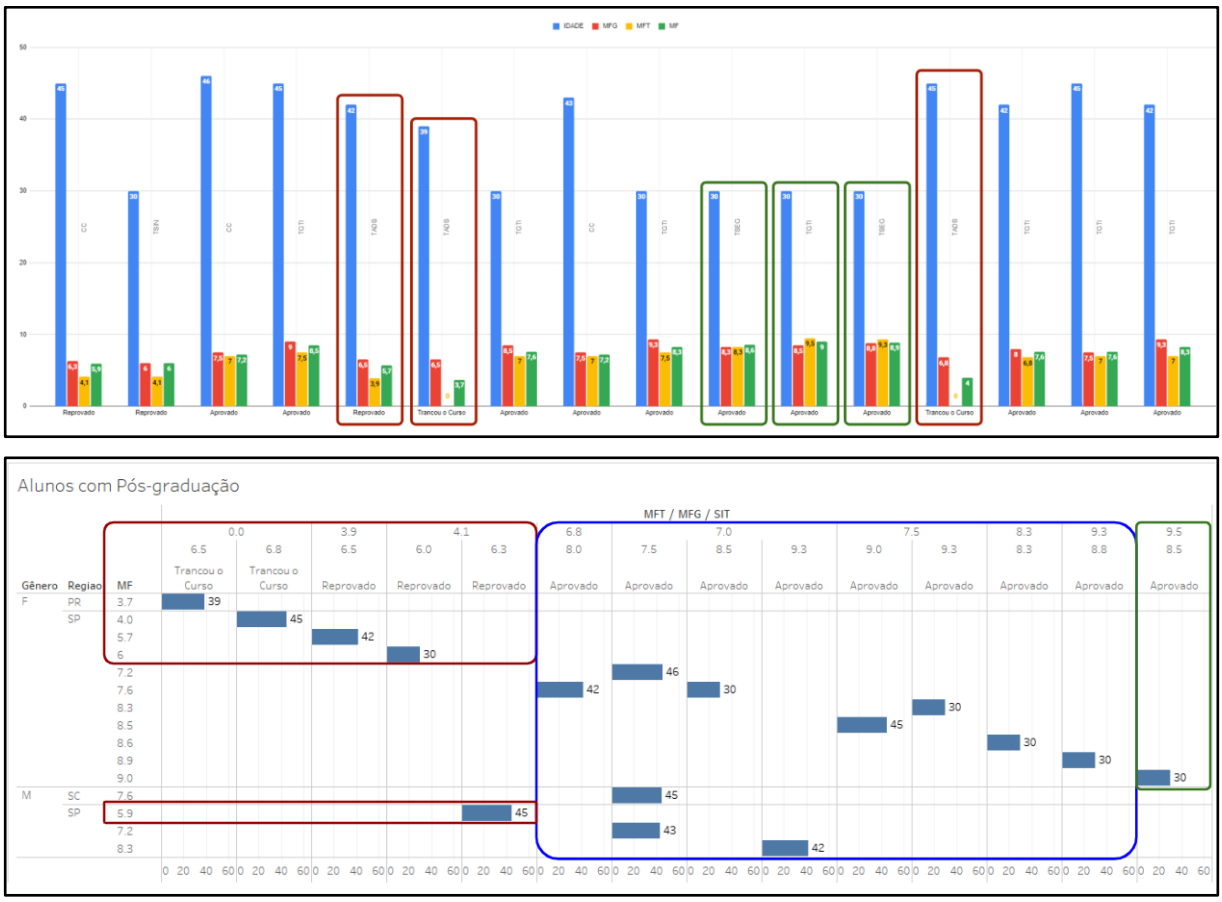
Fonte: Autora (2023)

É possível observar os cursos de graduação nos quais os alunos de pós-graduação se titularam, bem como ao estado ao qual eles pertencem. Assim sendo, identifica-se a origem e características de cada aluno matriculado. Neste grupo fica evidente que a maioria dos egressos frequentou o curso de TGTI em suas graduações, o que nos leva a entender que os alunos ao optarem por uma especialização de Governança em TI buscam seguir, ou manter, uma carreira com o foco específico em gestão da tecnologia, dentro da área principal de Informática, eles ocupam 43.75%, com 7 alunos.

Do ponto de vista de região específica, São Paulo ainda supera os demais estados com 14 alunos, representando 87.5%, o que leva a entender que este estado possui um perfil de pessoas que estão buscando elevar o nível de sua escolaridade, por melhorar seu desempenho profissional, ou buscar maiores cargos e salários.

Na Figura 44 observa-se como cada aluno já pós-graduado performou, analisando-se os atributos média final das disciplinas gerenciais (MFG), média final das disciplinas técnicas (MFT) e média final total das disciplinas (MF).

Figura 44: Correlação entre os atributos média final das disciplinas gerenciais (MFG), média final das disciplinas técnicas (MFT) e média final total das disciplinas (MF)

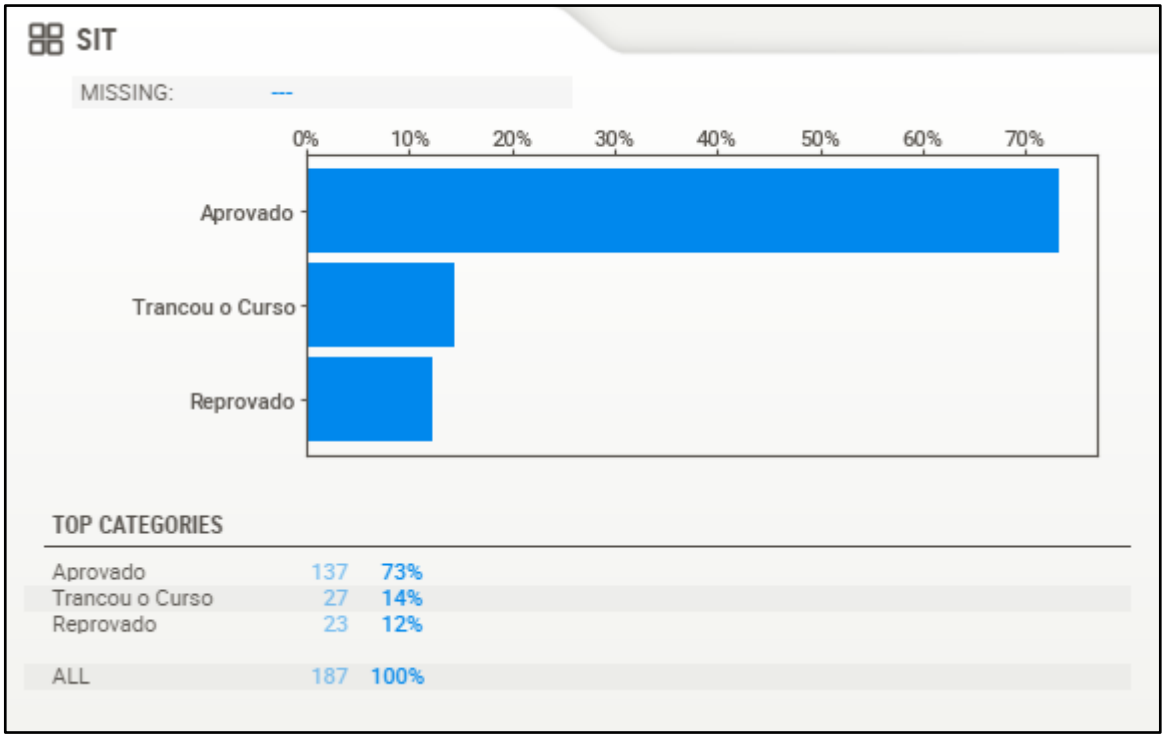


Fonte: Autora (2023)

Verifica-se uma aluna do gênero feminino com idade de 30 anos, egressa do curso TGTI (Tecnologia em Gestão de Tecnologia da Informação) e anteriormente cursou o ensino médio na modalidade normal. Neste curso de Governança em TI esta aluna obteve maior êxito nas disciplinas técnicas, obtendo uma média 9,5 e 8,5 nas disciplinas relacionadas a matérias gerenciais, conquistando a média final 9,0 ao concluir o curso. Em contrapartida, o aluno que possui o menor desempenho foi também do gênero feminino, porém, com 39 anos egressa do curso de TADS (Tecnologia em Análise e Desenvolvimento de Sistemas), cursou o ensino médio na modalidade normal, obteve a média 6,5 nas disciplinas relacionadas a matérias gerenciais e não cursou as disciplinas de cunho técnico, pois trancou o curso ao final da sexta disciplina.

Na Figura 45 é apresentado o resultado do atributo ‘situação’ dos alunos (aprovado, reprovado e trancou o curso).

Figura 45: Distribuição dos alunos no atributo ‘situação’

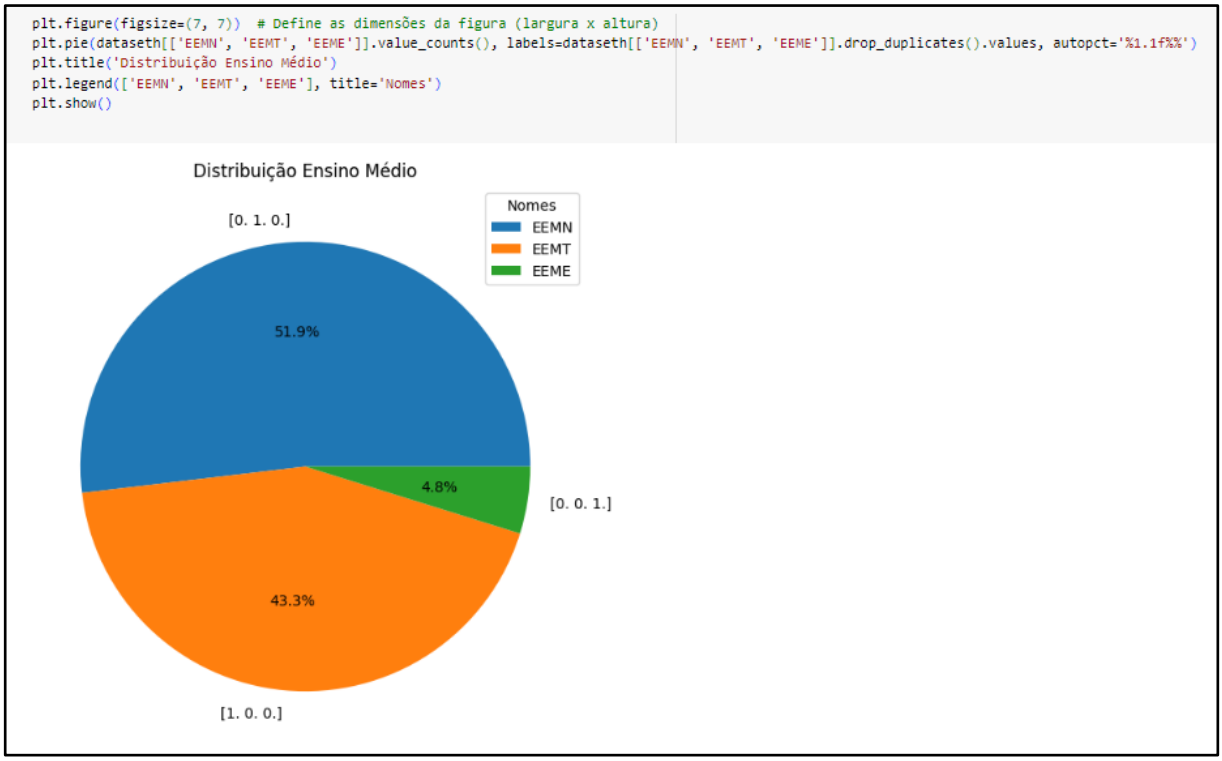


Fonte: Autora (2023)

O atributo ‘situação’ representa o status do aluno ao desfecho do curso de pós-graduação. Observa-se que a maioria dos alunos obteve êxito em seus estudos, sendo aprovados. Dos 187 alunos analisados, 137 (aproximadamente 73%) do total, alcançaram a conclusão satisfatória do curso. Por outro lado, 27 alunos (cerca de 14%) optaram por interromper o curso (trancou o curso), enquanto 23 alunos (aproximadamente 12%) foram reprovados.

Os alunos do curso de Governança de TI cursaram três tipos de ensino médio, quais sejam: EEMN (Escolaridade Ensino Médio Normal), EEMT (Escolaridade Ensino Médio Técnico) e EEME (Escolaridade Ensino Médio EJA). A Figura 46 apresenta os resultados totais de alunos dos cursos de ensino médio (Normal, Técnico e EJA).

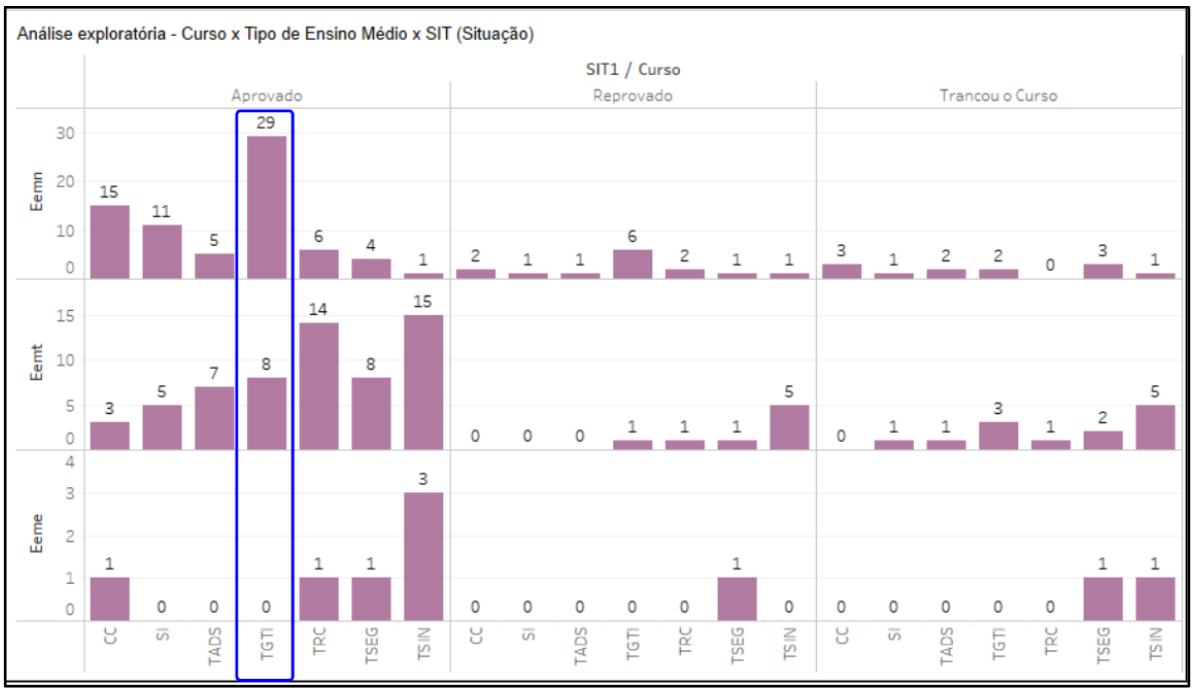
Figura 46: Distribuição dos alunos Ensino Médio (EEMN, EEMT e EJA)



Fonte: Autora (2023)

A análise da distribuição dos dados no conjunto mostra que os alunos apresentam diferentes formações de Ensino Médio. Cerca de 51,9% dos alunos cursaram o Ensino Médio Normal (EEMN), enquanto aproximadamente 43,3% cursaram o Ensino Médio Técnico (EEMT). Por fim, observa-se que uma parcela de 4,8% dos alunos teve sua formação em Ensino Médio na modalidade de Educação de Jovens e Adultos (EEME). Esses resultados revelam a diversidade de trajetórias educacionais dos alunos presentes no conjunto de dados, com uma significativa proporção tendo cursado o Ensino Médio Normal ou Técnico. A presença de alunos com formação em Educação de Jovens e Adultos também indica a inclusão de diferentes perfis de estudantes na análise. Essas informações podem fornecer percepções importantes sobre o perfil educacional dos alunos e auxiliar na compreensão de possíveis correlações com outras variáveis do conjunto de dados.

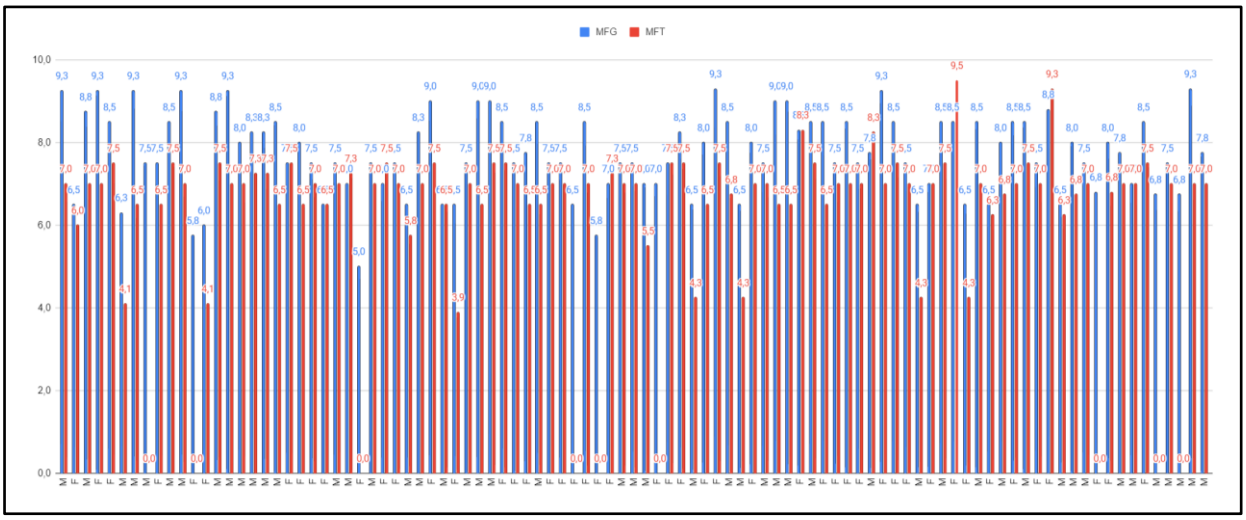
Figura 47: Distribuição dos alunos no atributo ‘situação’ alunos do ensino médio (Normal, Técnico e EJA) por curso de graduação



Fonte: Autora (2023)

É possível notar que dos 187 alunos analisados, 49 cursaram a graduação TGTI (Tecnologia em Gestão de Tecnologia da Informação), isto refere-se a 26,2% em relação aos demais cursos, sendo que destes, 19,3% (36 alunos) fizeram o EEMN (Escolaridade Ensino Médio Normal) e performando melhor, obtendo o maior índice de aprovação no curso de pós-graduação em Governança de TI. Na Figura 48 observa-se estes alunos vistos sob a análise das médias das disciplinas gerenciais e técnicas. Na figura 48 são apresentados os resultados do atributo EEMN em análise das Médias das disciplinas gerenciais e técnicas.

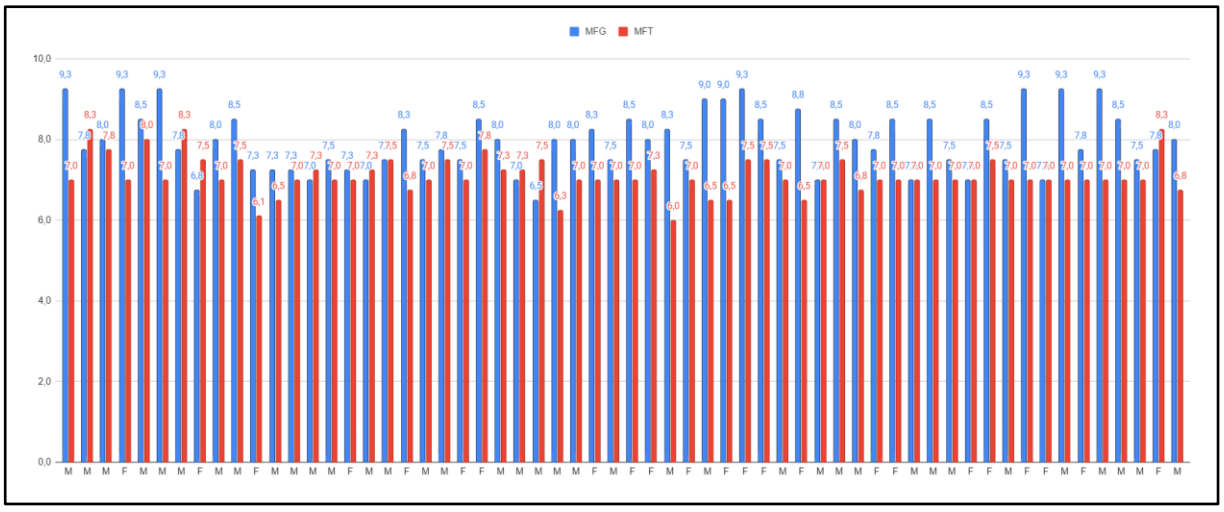
Figura 48: Resultados do atributo EEMN em análise das Médias das disciplinas



Fonte: Autora (2023)

Nota-se que a média das disciplinas gerenciais possui destaque, uma vez que de 97 alunos que cursaram esta modalidade, 42 alunos (25 homens e 17 mulheres) obtiveram notas acima de 8,0 nas disciplinas gerenciais e abaixo de 7,9 nas disciplinas técnicas o que corresponde a 43,3% desta população. Em relação ao inverso, analisando-se os alunos que obtiveram notas acima de 8,0 nas disciplinas técnicas e abaixo de 7,9 nas disciplinas gerenciais chegamos ao número de 22 alunos (11 homens e 11 mulheres), que equivale a 22,7% dos alunos estudados. O que nos mostra que os alunos egressos do ensino médio normal performam melhor nas disciplinas gerenciais. A Figura 49 apresenta o desempenho dos alunos do ensino médio técnico, comparando as disciplinas gerenciais e técnicas.

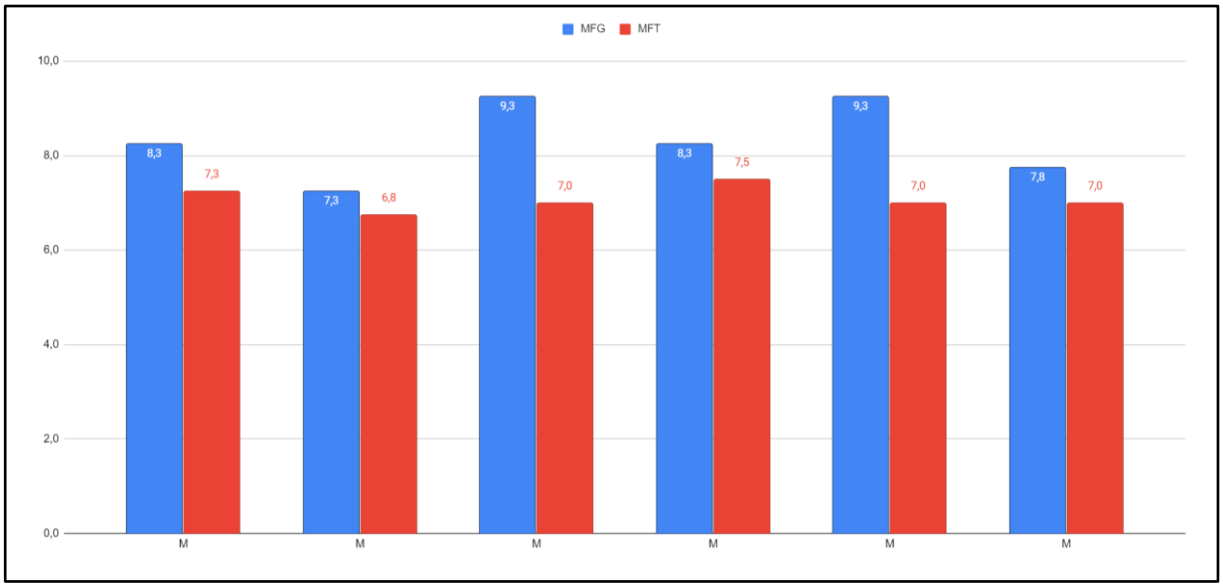
Figura 49: Resultados do atributo EEMT em análise das Médias das disciplinas



Fonte: Autora (2023)

Neste grupo a média das disciplinas gerenciais e técnicas é mais balanceada, pois, de 60 alunos aprovados que cursaram esta modalidade técnica, 51 alunos (32 homens e 19 mulheres) obtiveram notas acima de 7,0 em ambas as disciplinas gerenciais e técnicas o que corresponde a 85% desta população. A Figura 50 apresenta o desempenho dos alunos do ensino médio EJA.

Figura 50: Resultados do atributo EEJA em análise das Médias das disciplinas



Fonte: Autora (2023)

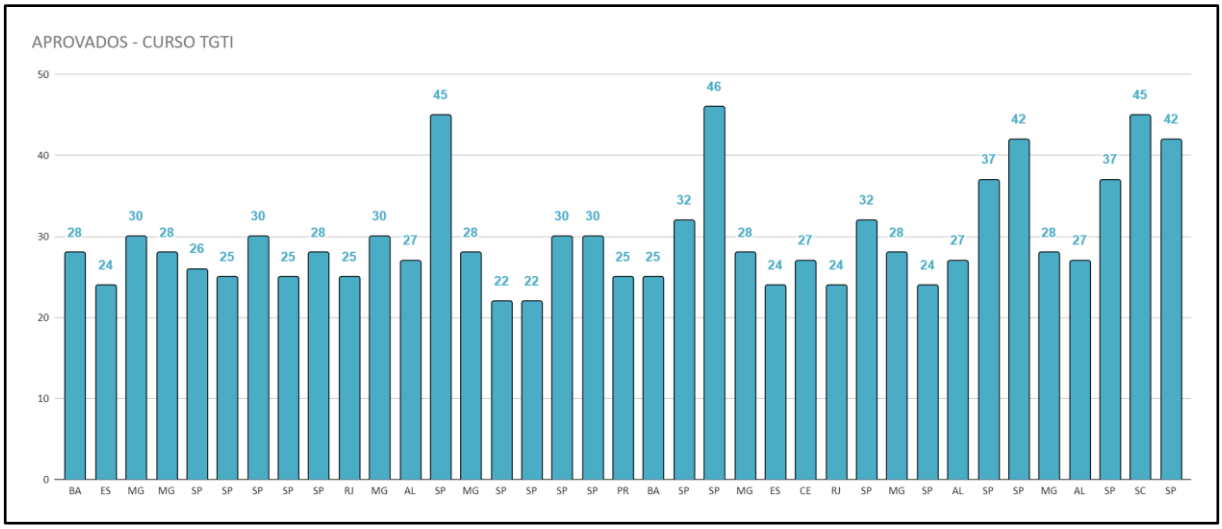
Observa-se um grupo menor, porém, com médias maiores para as disciplinas de gerenciais, posto que obtiveram médias entre 7,8 e 9,3 nas seis disciplinas gerenciais, e de 6,8 a 7,5 nas disciplinas técnicas.

Os resultados das médias obtidas pelos alunos egressos de diferentes modalidades de Ensino Médio indicam que o ensino médio normal, ensino médio técnico e ensino médio EJA pós-graduandos no curso de Governança em TI obtiveram notas médias mais elevadas nas disciplinas gerenciais do curso, e não nas disciplinas técnicas.

O desempenho obtido pelos alunos egressos de cursos de graduação mencionados anteriormente, em especial o curso TGTI (Tecnologia em Gestão de Tecnologia da Informação) apresenta maior destaque na pós-graduação em

Governança de TI. Analisando-se este público expressivo, a Figura 51 mostra a população de alunos que tiveram melhor êxito, sendo analisados por idades.

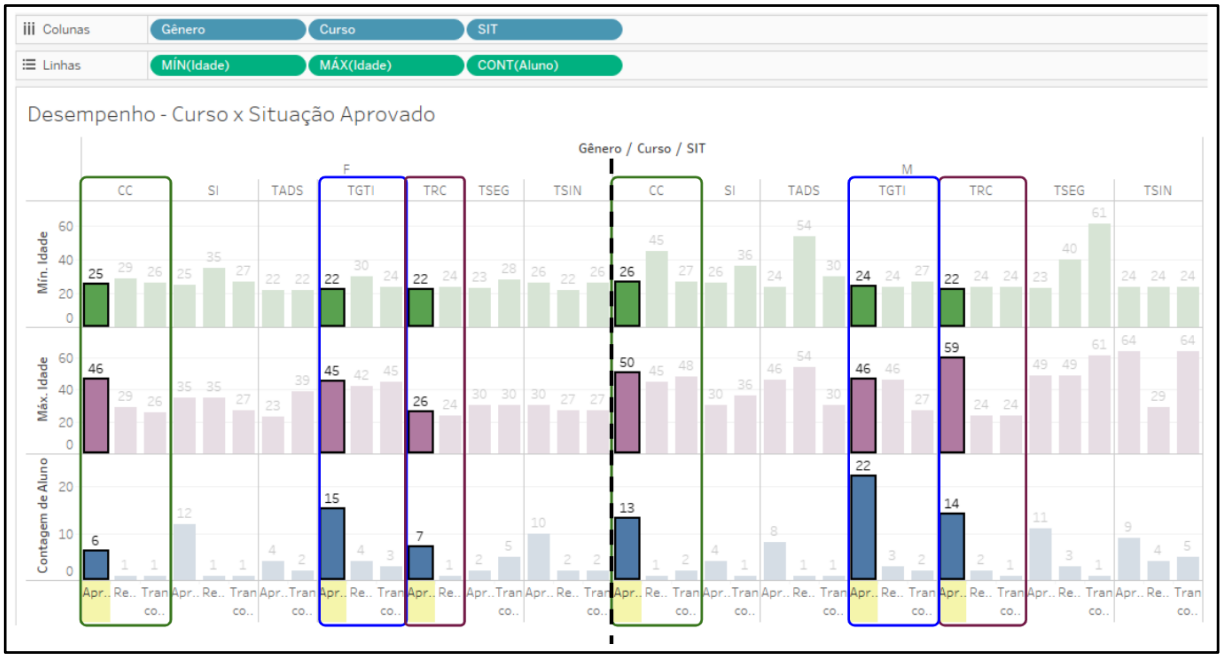
Figura 51: Correlação entre os atributos ‘Aprovados’, ‘curso de graduação TGTI’ e ‘idades’



Fonte: Autora (2023)

A população de 38 aprovados neste curso possui idades que variam de 22 anos a 46 anos com a média de 30 anos. Este grupo demonstrou melhor desempenho mediante os demais cursos, o que leva a entender que alunos com a média de 30 anos em busca de especialização terão maior probabilidade de melhor desempenho caso venham ingressar em um curso de especialização em Governança de TI. Na Figura 52 é exposta a correlação entre o atributo ‘situação’ e o curso de graduação do aluno.

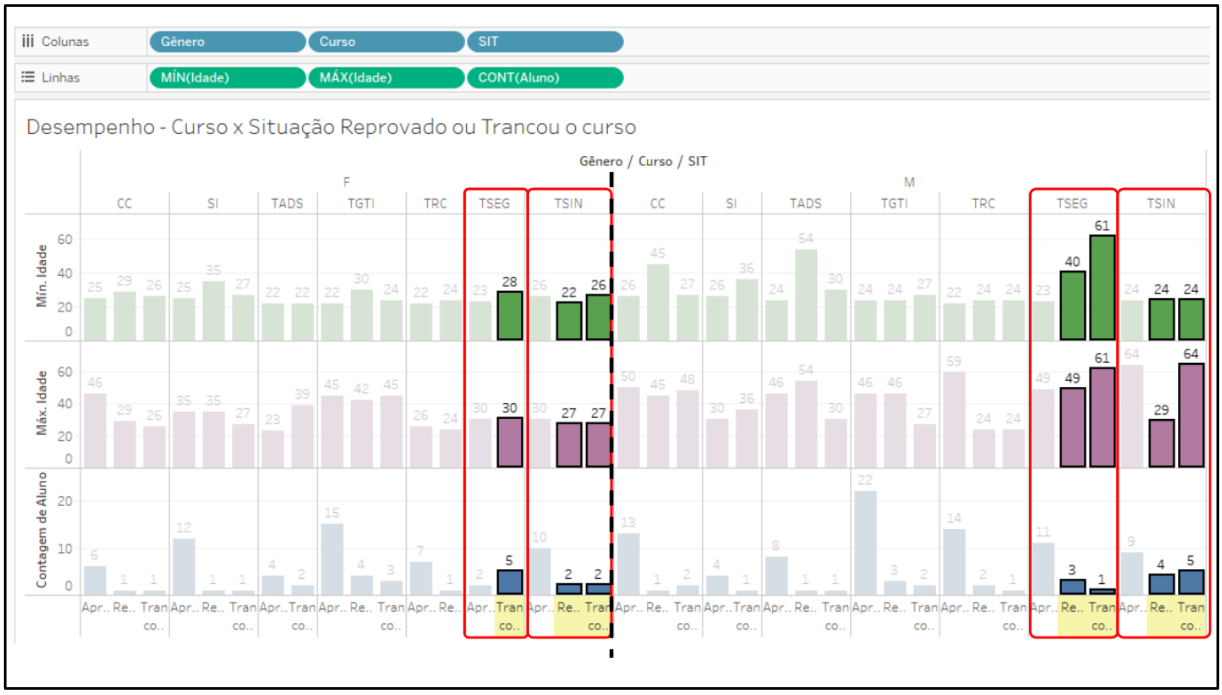
Figura 52: Correlação entre os atributos ‘Aprovados’ e ‘curso de graduação’



Fonte: Autora (2023)

O curso de TGTI obteve o maior número de aprovação da população estudada, com 37 alunos aprovados de um total de 137 alunos aprovados no curso de pós-graduação em Governança de TI, o que equivale a 27% dos alunos analisados desta base de dados, distribuídos com idades entre 22 e 46 anos com média aproximada de 30 anos. Em segundo lugar está o curso de TRC com 21 alunos aprovados, que representa 15,3% dos aprovados, distribuídos com idades entre 22 e 59 anos com média aproximada de 26 anos, porém com maior massa de alunos de idade entre 22 e 25 anos, e por fim o curso de CC que aparece com 19 alunos aprovados dentre este total, distribuídos com idades entre 25 e 50 anos com média aproximada de 35 anos. Alunos mais velhos comparados aos demais alunos analisados, o que equivale aproximadamente a 13,9% de aprovação. Na Figura 53 são expostos os resultados a partir dos cursos que obtiveram o menor desempenho.

Figura 53: Correlação entre os atributos Reprovados ou Trancou o curso, ‘curso de graduação’



Fonte: Autora (2023)

Considerando a situação oposta, os cursos que tiveram o menor desempenho analisando um total de 50 alunos reprovados ou que trancaram o curso, foram os cursos de TSIN (Tecnologia em Sistemas para Internet), com 6 alunos reprovados e 7 alunos que trancaram o curso, obtendo total de 13 alunos, que representa 26% de alunos que performaram de maneira insatisfatória, e o curso de TSEG (Tecnologia em Segurança da Informação), com 3 alunos reprovados e 6 alunos que trancaram o curso, obtendo total de 9 alunos, que representa 18% de alunos que não conquistaram um desempenho satisfatório.

A partir dos resultados é possível concluir que o desempenho de alunos provenientes do Ensino Médio (modalidades EEMN - ensino médio normal, EEMT - ensino médio técnico e EEME - ensino médio EJA), os alunos obtiveram os melhores resultados nas disciplinas gerenciais do curso. Destes se destacaram alunos que após a conclusão do EEMN (ensino médio normal), cursaram a graduação de TGTI (Tecnologia em Gestão da Tecnologia da Informação), com 29 aprovados com este perfil, com idades entre 22 e 46 anos, e média de 30 anos.

Observando-se apenas os cursos de graduação, o curso de TGTI obteve o maior número de aprovação da população estudada, com 37 alunos aprovados de um

total de 137 alunos aprovados no curso de pós-graduação em Governança de TI, o que equivale a 27% dos alunos analisados desta base de dados, distribuídos com idades entre 22 e 46 anos com média aproximada de 30 anos. Em segundo lugar está o curso de TRC com 21 alunos aprovados, que representa 15,3% dos aprovados, distribuídos com idades entre 22 e 59 anos com média aproximada de 26 anos, porém com maior massa de alunos de idade entre 22 e 25 anos. Por fim, o curso de CC que aparece com 19 alunos aprovados dentre este total, distribuídos com idades entre 25 e 50 anos com média aproximada de 35 anos. Alunos mais velhos comparados aos demais alunos analisados, o que equivale aproximadamente a 13,9% de aprovação.

Em contrapartida, os cursos que tiveram o menor desempenho analisando-se um total de 50 alunos reprovados ou que trancaram o curso, foram os cursos de TSIN (Tecnologia em Sistemas para Internet), com 6 alunos reprovados e 7 alunos que trancaram o curso, obtendo total de 13 alunos, que representa 26% de alunos que performaram de maneira insatisfatória. O curso de TSEG (Tecnologia em Segurança da Informação), com 3 alunos reprovados e 6 alunos que trancaram o curso, com um total de 9 alunos, que representa 18% de alunos que não conquistaram um desempenho satisfatório.

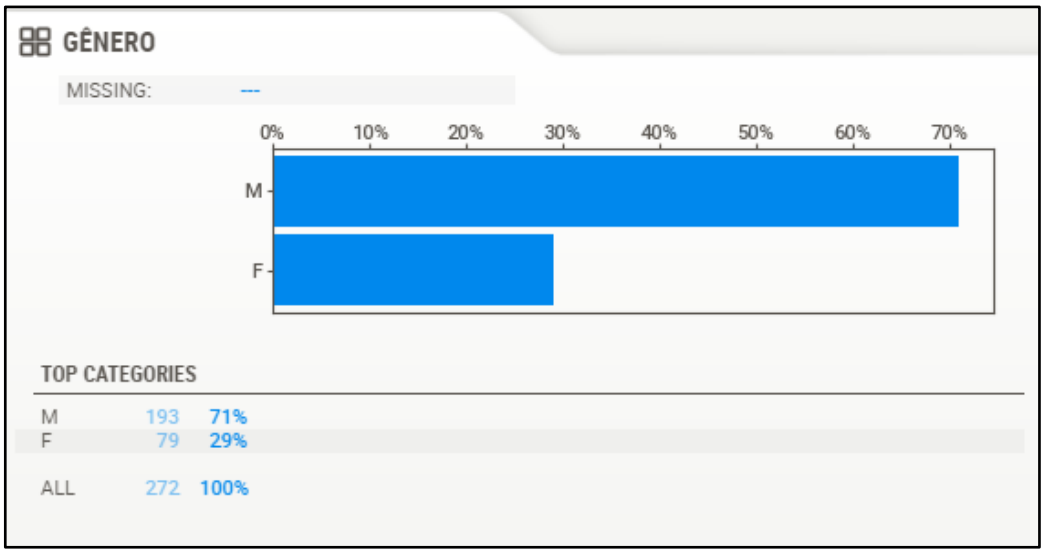
Perfis a serem considerados

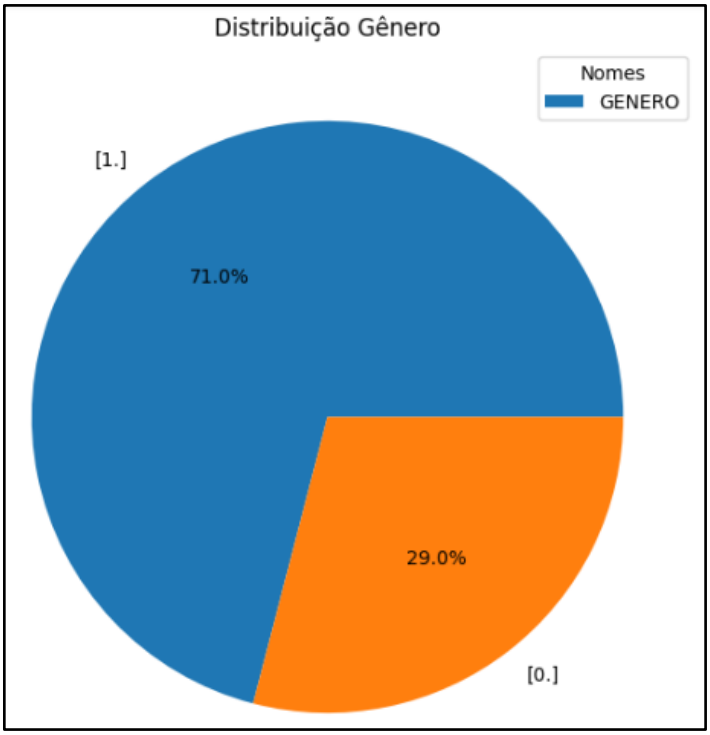
Em conclusão, alunos que realizaram o EEMN (ensino médio normal), cursaram a graduação de TGTI (Tecnologia em Gestão da Tecnologia da Informação), com idade média de 30 anos, do gênero feminino e que já realizaram um curso de especialização tendem a performar melhor que outros perfis. O oposto se dá quano aos alunos que realizaram o EEMT (ensino médio técnico), do curso de graduação TSIN (Tecnologia em Sistemas para Internet), com idade média de 29 anos do gênero masculino, que não possuem previamente uma pós-graduação.

4.2 Resultados Curso Data Science

Na Figura 54 são apresentadas as análises realizadas por meio da ferramenta *Google Colab* para o curso de Data Science, com a distribuição proporcional de alunos com base no atributo gênero.

Figura 54: Percentual de alunos por categoria/atributo - Gênero

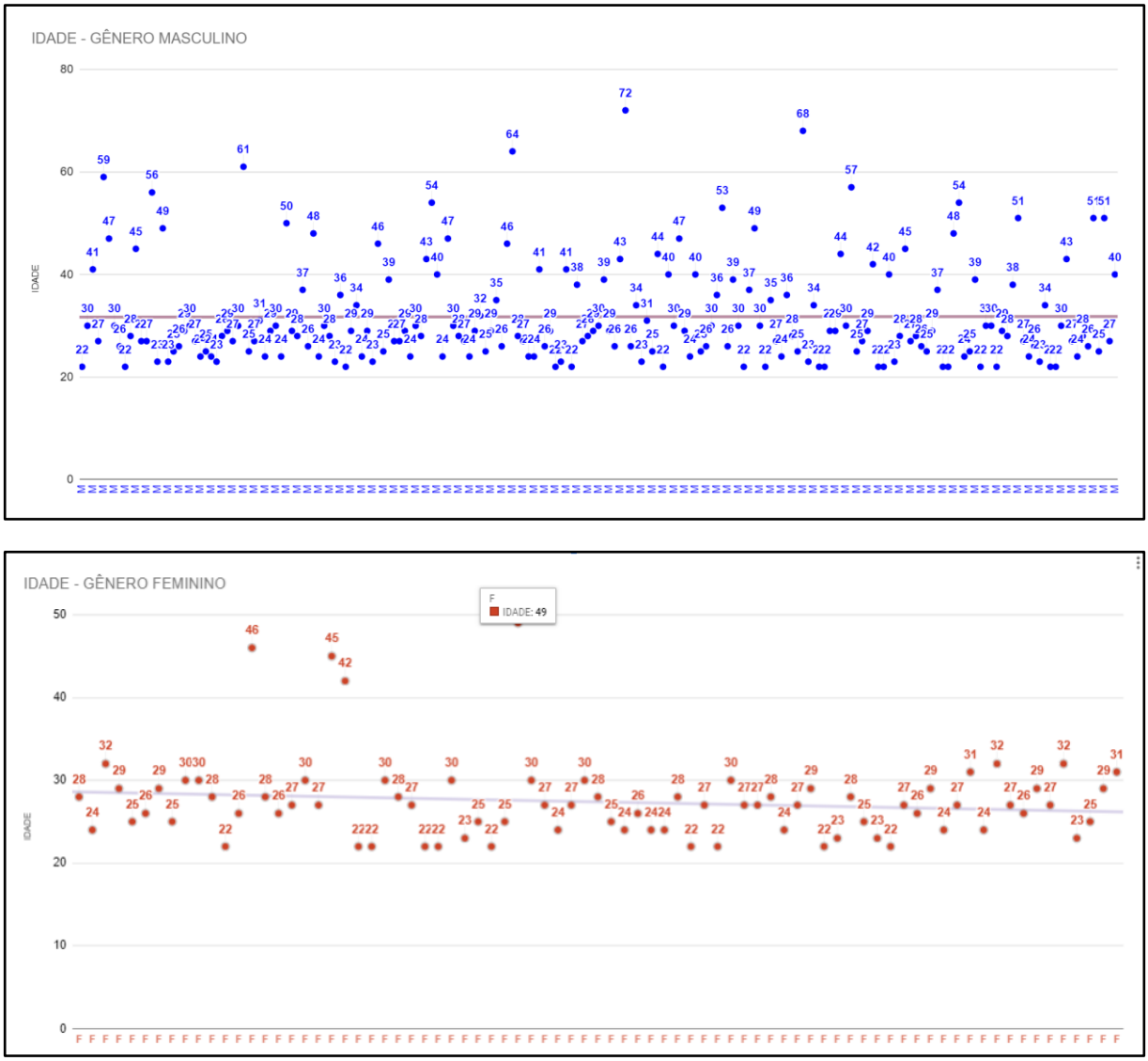




Fonte: Autora - Google Colab (2023)

A distribuição de alunos por gênero evidencia uma prevalência masculina. Dos 272 alunos no conjunto de dados, 193 são do sexo masculino (1), o que representa aproximadamente 71% da população, enquanto 79 são do sexo feminino (0), correspondendo a cerca de 29%. A distribuição mais detalhada da idade dos alunos(as) por gênero é exposta na Figura 55.

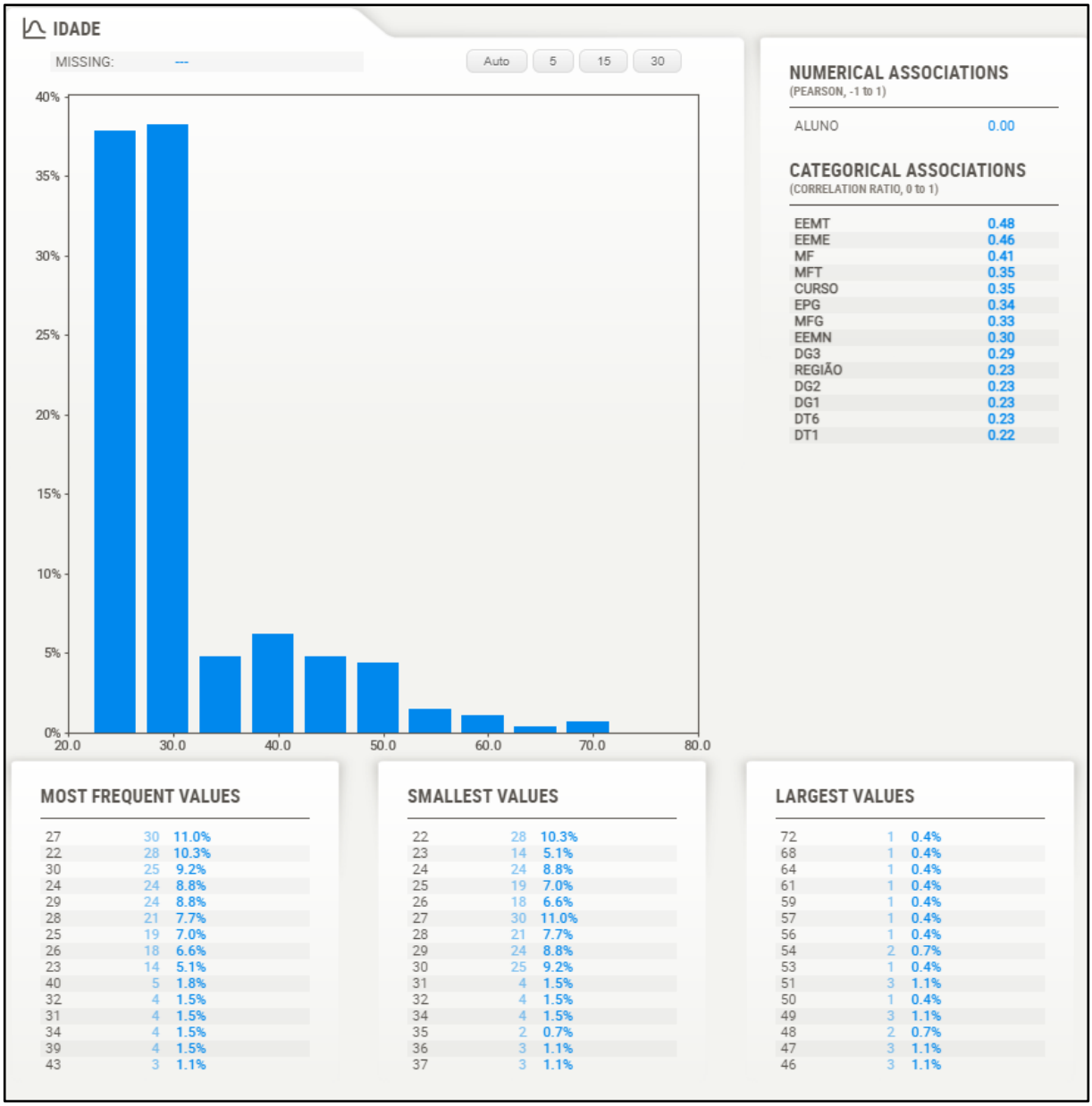
Figura 55: Distribuição dos alunos no atributo ‘gênero’



Fonte: Autora (2023)

A distribuição no grupo masculino ocorre em idades de 22 a 72 anos com a média de 32 anos de cada estudante. Considerando-se o grupo feminino, a distribuição ocorre em idades que variam entre de 22 a 49 anos, com a média de 27 anos. A Figura 56 expõe a distribuição dos alunos pelo atributo ‘idade’.

Figura 56: Distribuição dos alunos por atributo “idade”

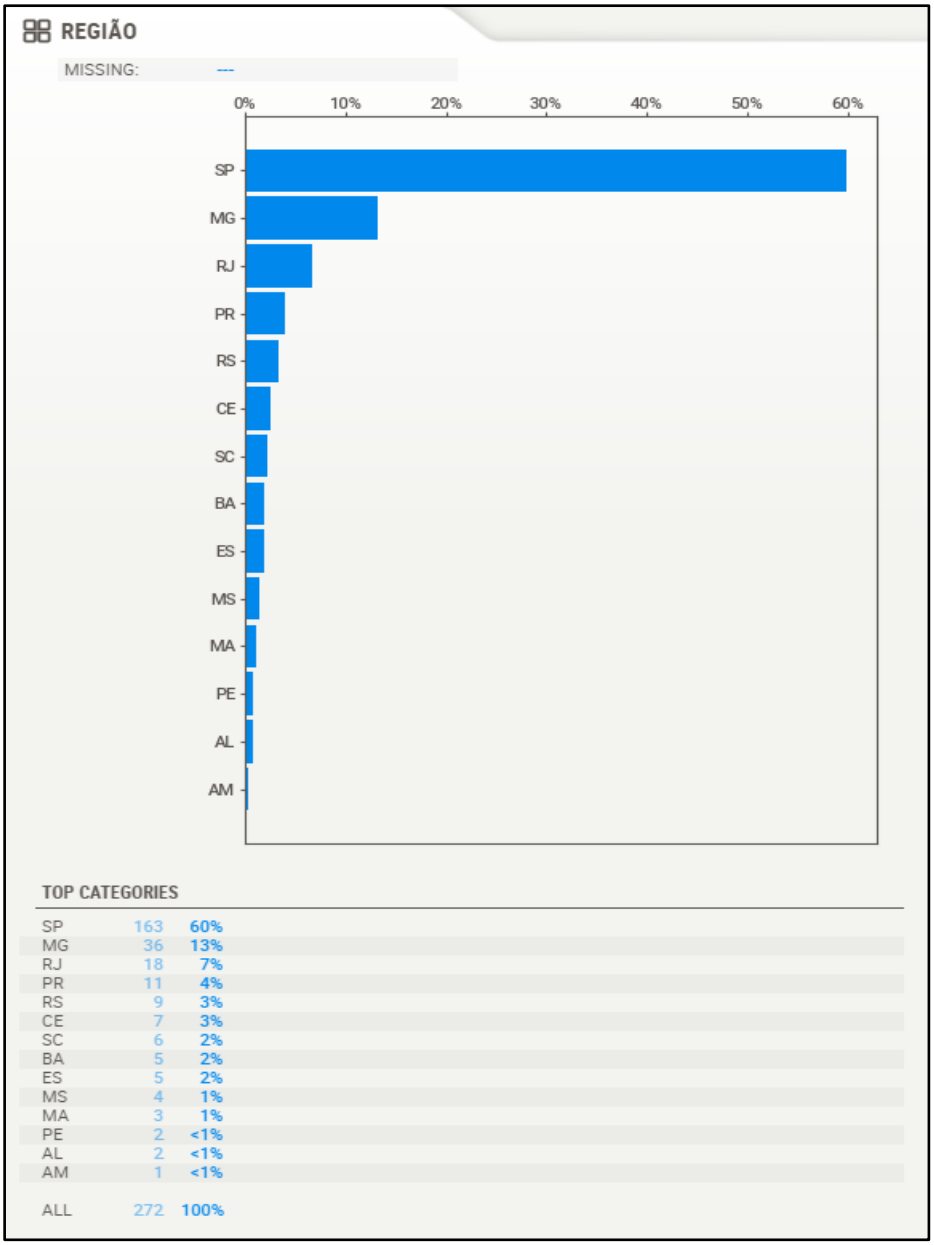


Fonte: Autora (2023)

Os resultados revelam que a faixa etária mais comum entre os alunos é de 22 a 30 anos. O que mostra uma população um pouco mais jovem inserida neste tipo de curso de especialização. Em um total de 272 alunos, 11% dos alunos têm 27 anos, 10,3% têm 22 anos, 9,2% têm 30 anos, 8,8% têm 24 anos. Evidencia-se um nicho de alunos iniciando suas carreiras voltados à tendência do estudo em Data Science, buscando especialização nas novas abordagens de estudo de Ciência e análise de dados. Já em último colocado neste grupo verificaram-se alunos acima de 43 anos, que representam cerca de 1% da população do curso.

Na Figura 44 são apresentados os resultados do atributo ‘região’ (estado de domicílio no Brasil).

Figura 57: Percentual de alunos por atributo “região”



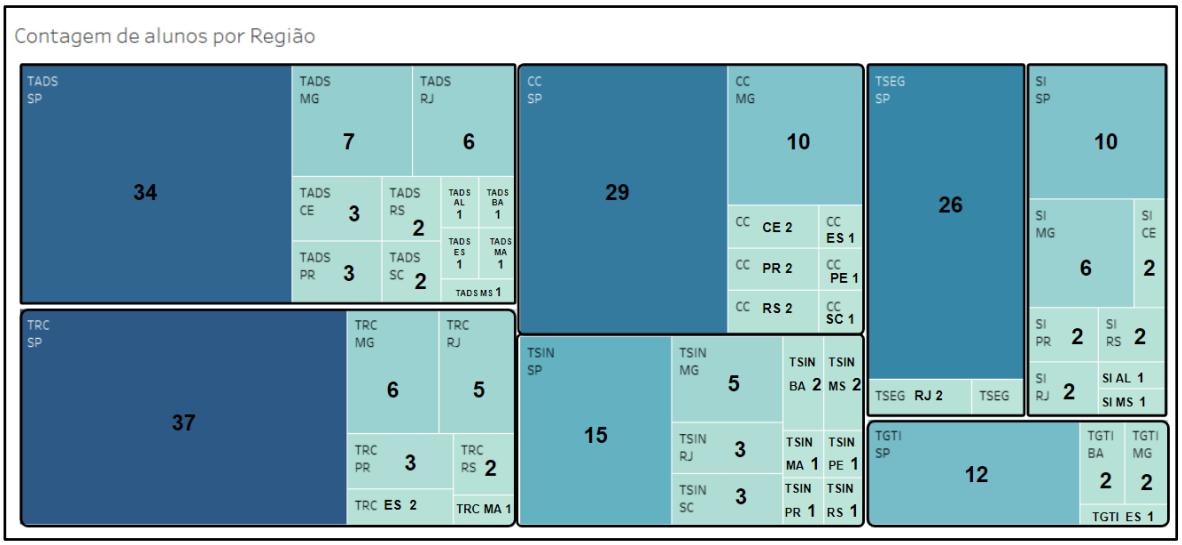
Fonte: Autora (2023)

É possível notar que a grande massa dos alunos é residente da região São Paulo, com 163 alunos, representando aproximadamente 60% do total de matrículas na pós-graduação em Data Science. Em segundo lugar está o estado de Minas Gerais com 36 alunos, correspondendo a 13% e em terceiro lugar o estado do Rio de Janeiro com 18 alunos, representando 7%, em quarto lugar o estado de Paraná com 11

alunos, representando 4% e em quinto lugar o estado do Rio grande do Sul com 9 alunos, aproximadamente 3% do total analisado.

A Figura 58 expõe os resultados dos atributos ‘região’ e ‘curso de graduação’.

Figura 58: Correlação entre os atributos “região” e “curso”

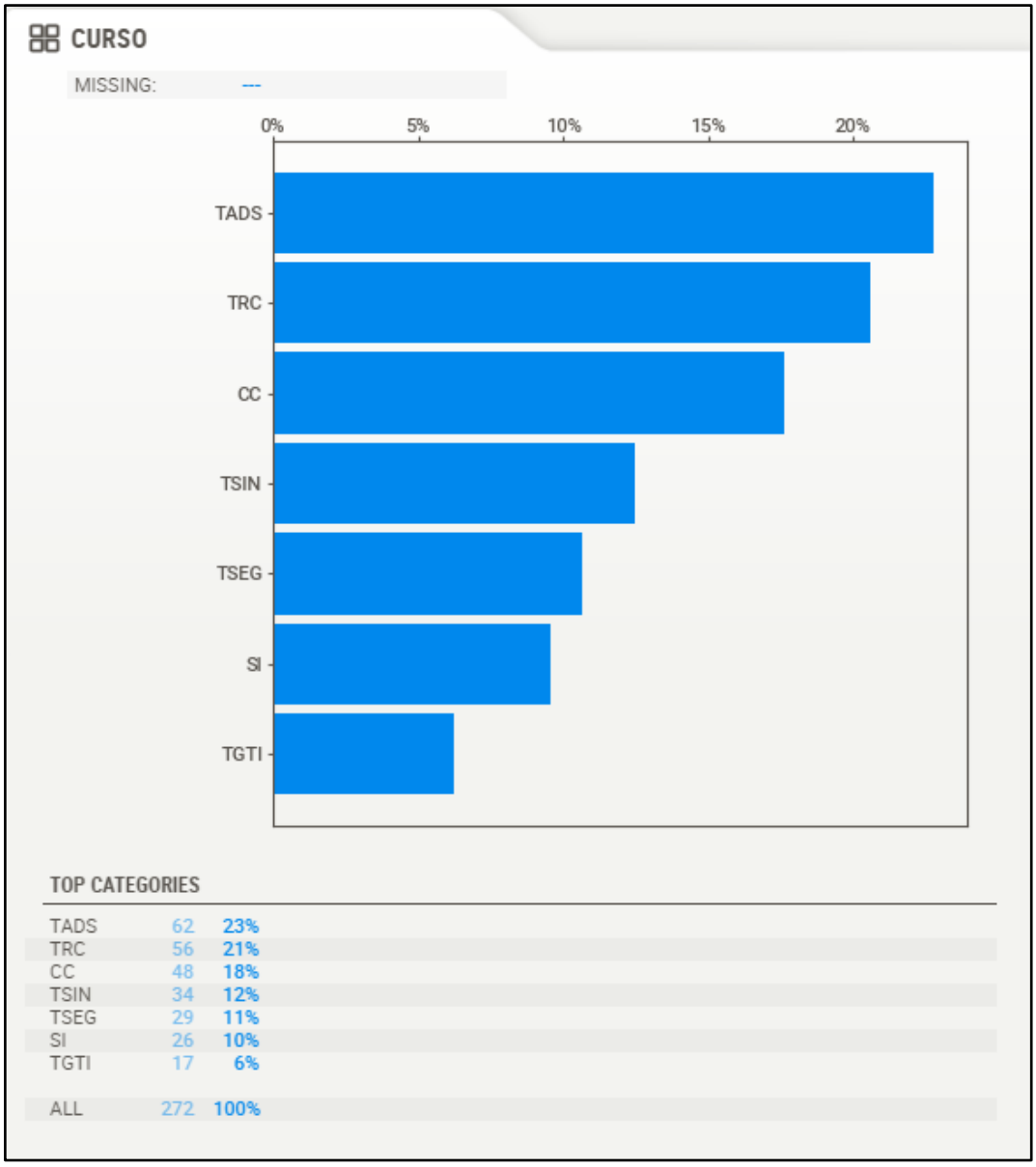


Fonte: Autora (2023)

Na Figura 58 foi apresentada a distribuição de alunos por curso e região do país, onde verifica-se novamente que a maior parte dos alunos é residente do estado de São Paulo, mediante aos demais estados neste atributo ‘Região’. Analisando separadamente cada curso, nota-se que o curso de TRC é o que possui a maior concentração de alunos, em São Paulo são 37 alunos dos 163 analisados desta região, o que equivale a 22,7%, tendo o segundo lugar o curso de TSEG com 20 alunos representando 12,2% e por fim, TRC com 15 alunos obtendo a 9,2% dos alunos analisados nesta população. Diferente do curso de Governança em TI o curso de Data Science obteve a menor adesão de egressos do curso de TGTI, com apenas 12 alunos no estado de São Paulo, que representa 7,3%.

Na Figura 59 são apresentados os resultados da correlação entre os atributos ‘idade’ e ‘curso de graduação’

Figura 59: Distribuição de alunos no atributo “curso”



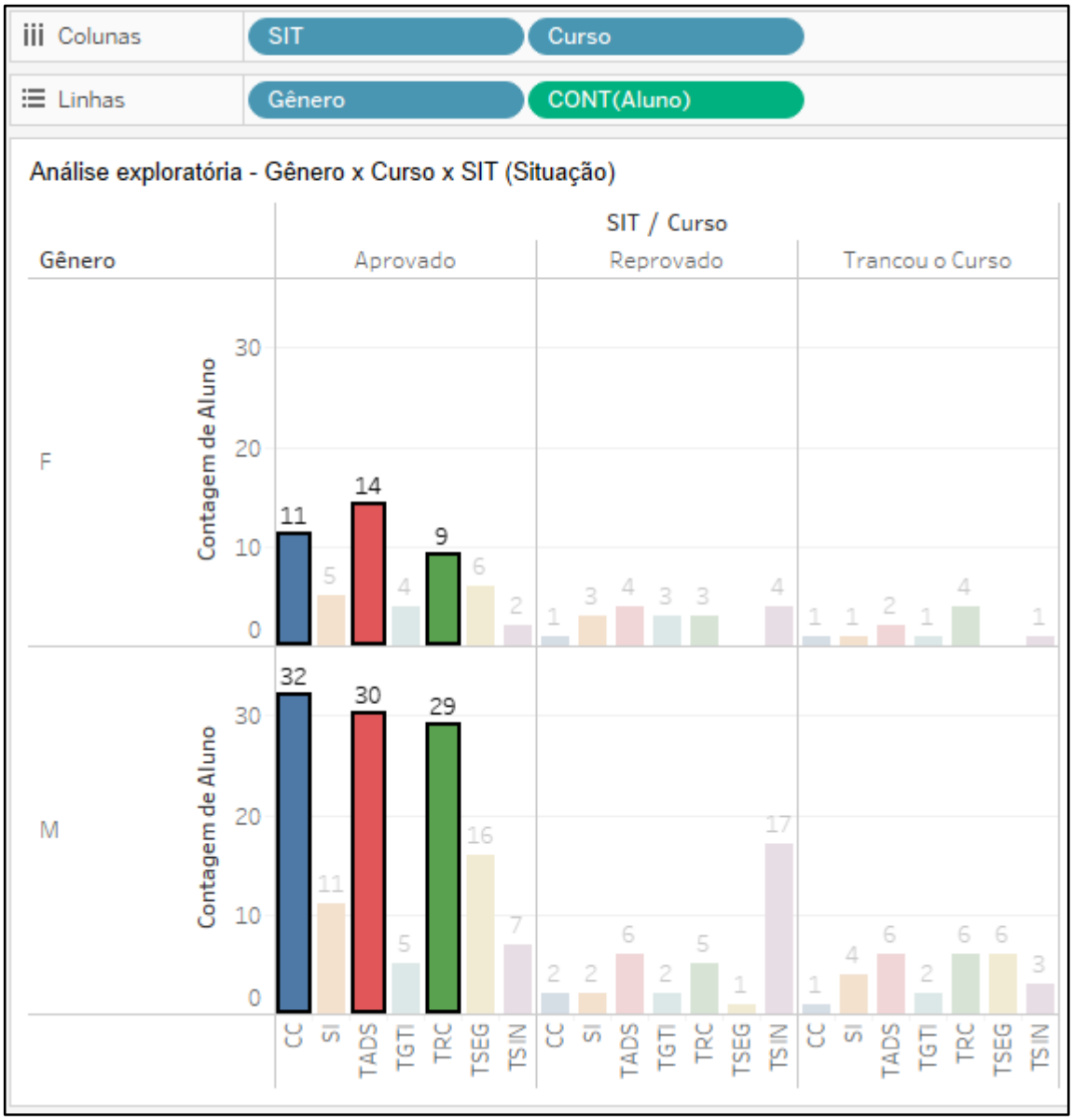
Fonte: Autora (2023)

Na Figura 59 é possível visualizar as porcentagens referentes à categoria “Cursos”, que mostra o número de alunos egressos do curso de TADS. Diferente da análise feita ao curso de pós-graduação analisado no capítulo anterior (4.1 Governança em TI), o curso de Data Science apresenta-se com maior número de alunos, são 62 alunos em 272, o que corresponde a 23%. Seguido do curso TRC, com 56 alunos, representando 21% e pelo curso CC, com 48 alunos, somando 18% do total estudado.

A participação dos egressos do curso de TGTI para este curso de pós-graduação é a menor, apenas 17 alunos em 272, o que corresponde a 6% do total de alunos analisados. A Figura 60 a seguir mostra a de forma cruzada os atributos Gênero, Curso e SIT (Situação).

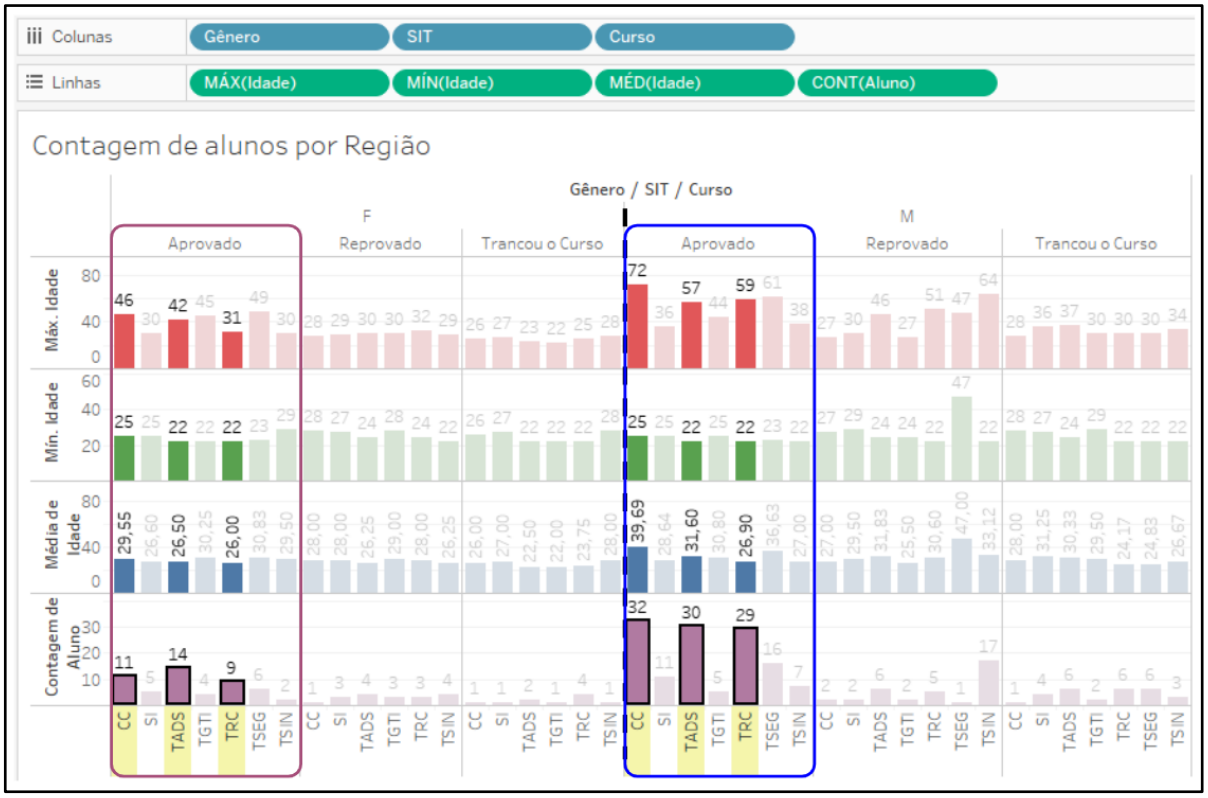
A Figura 60 apresenta a relação entre os atributos “gênero”, “curso de graduação” e “situação”.

Figura 60: Correlação entre os atributos “gênero”, “curso de graduação” e “situação”



Fonte: Autora (2023)

Figura 61: Correlação entre os atributos “gênero”, “curso de graduação” e “situação”



Fonte: Autora - *Software: Tableau* (2023)

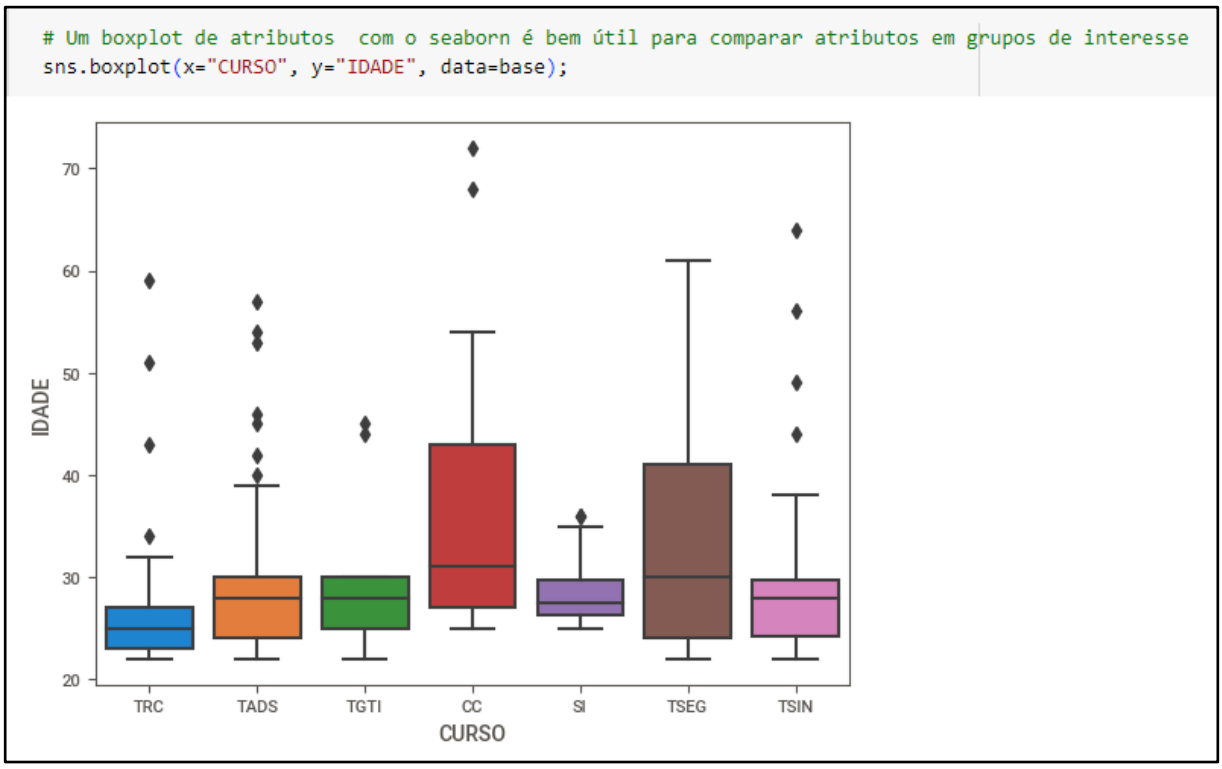
É possível perceber o melhor desempenho neste curso de pós-graduação para os alunos egressos do curso de TADS (Tecnologia em Análise e Desenvolvimento de sistemas), com 44 alunos no total (30 do gênero masculino e 14 do gênero feminino), que representam 16,2% de aprovados neste curso. Observando-se o atributo “gênero”, destacam-se os alunos do gênero masculino do curso de CC (Ciência da Computação), com 32 alunos aprovados, representando 11,8% de sucesso, com idades entre 25 e 72 anos. Tais resultados evidenciam grande quantidade de alunos formados no bacharelado voltando aos estudos após uma idade mais avançada, ou também alunos mais experientes em suas carreiras, buscando uma segunda especialização. O curso de TADS (Tecnologia em Análise e Desenvolvimento de sistemas), com 30 alunos aprovados, representando 11%, com idades entre 22 e 57 anos, e o curso de TRC (Tecnologia em Redes de Computadores), com 29 alunos, representando 10,7% de aprovação, com idades entre 22 e 59 anos, mostra um público masculino mais experiente na maioria dos cursos analisados.

Já ao público feminino destacou-se pelo curso de graduação TADS (Tecnologia em Análise e Desenvolvimento de sistemas), com 14 alunas aprovadas, entre 22 e 42

anos, representando 5,1%. Na sequência encontra-se o curso CC (Ciência da Computação), com 11 aprovações (4%), com idades entre 25 e 46 anos, que são mulheres mais velhas neste grupo. Em terceiro vem o curso de TRC (Tecnologia em Redes de Computadores) com 9 alunas aprovadas, correspondendo a 3,3%, com idades entre 22 e 31 anos, que mostra ser um grupo mais jovem neste perfil. Face aos resultados, identifica-se a maior procura de cursos mais técnicos para os alunos de perfil oriundo de cursos de graduação com desempenho em geral mais técnico.

Na Figura 62 apresenta-se, de forma mais detalhada, as correlações entre os atributos “idade”, “gênero” e “curso de graduação” e “situação”, permitindo uma visualização mais clara dos intervalos de idades dos cursos.

Figura 62: Correlação entre os atributos idade, gênero, curso de graduação e situação



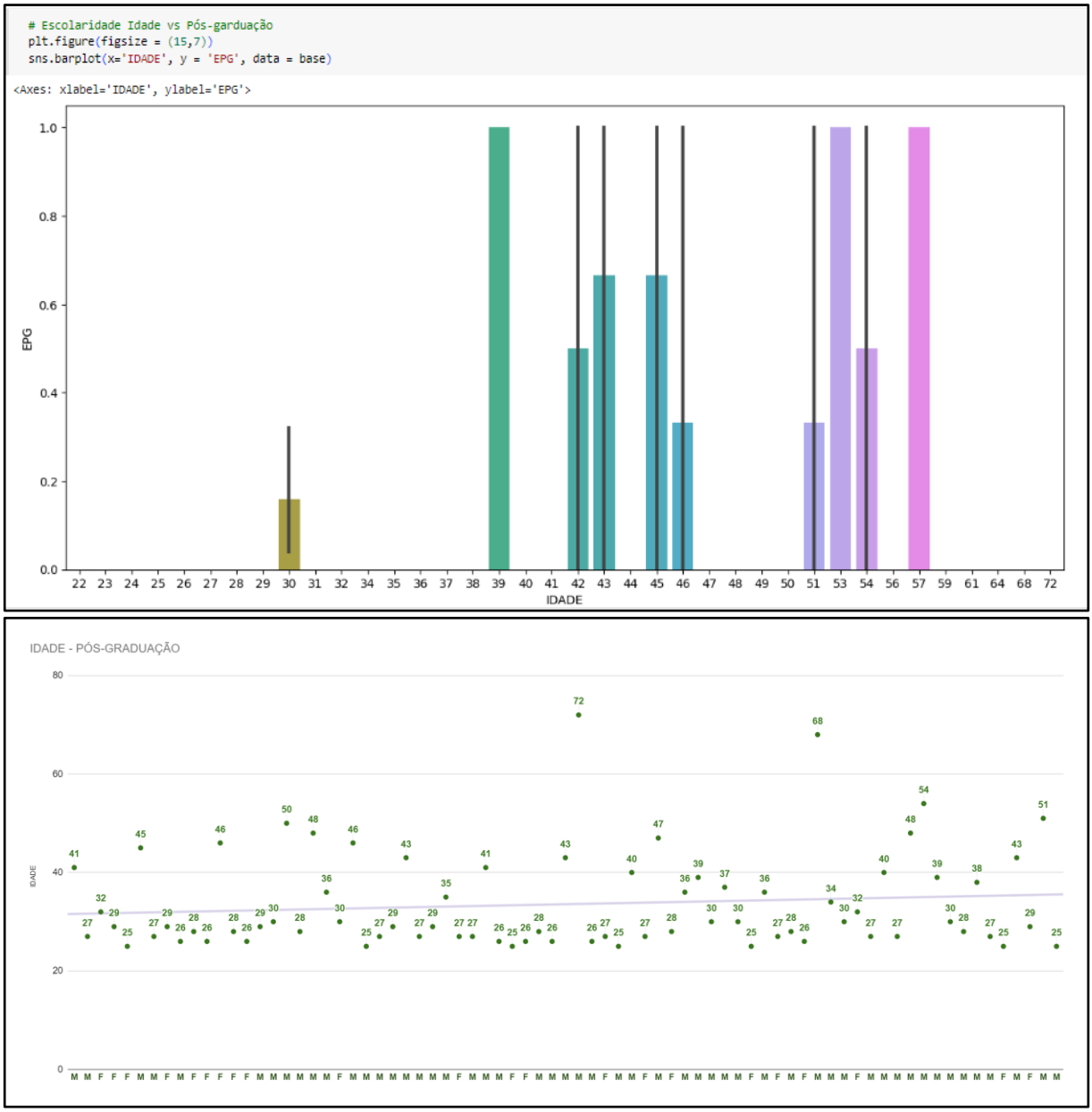
Fonte: Autora - Google Colab (2023)

Os resultados evidenciam uma tendência em relação à distribuição etária dos alunos em diferentes cursos de graduação. Os alunos mais jovens tendem a concentrar-se nos cursos de TRC (Tecnologia em Redes de Computadores), onde 50 alunos entre os 56 egressos deste curso, representa a faixa etária de entre 22 a 30 anos, o que equivale a 89,2% dos alunos matriculados vindos deste curso, no curso de TADS (Tecnologia em Análise e Desenvolvimento de Sistemas), de 62 alunos 49

possuem entre 22 e 30 anos, o que remete a 79% dos alunos. Em contrapartida, os alunos provenientes de cursos como CC (Ciência da Computação) 27 alunos de um total de 48, possuem entre 30 e 72 anos, que representa 56,25% desta população, e TSEG (Tecnologia em Segurança da Informação) 21 alunos de um total de 29, possuem entre 25 e 61 anos, que representa 72,4% deste grupo. Entendendo então que os alunos dos cursos como CC e TSEG abrangem um público mais velho em relação aos demais cursos.

Na Figura 63 é apresentada as correlações entre os atributos referentes aos alunos “pós-graduação” e “idade”.

Figura 63: Correlação entre os atributos ‘curso de Pós-graduação’ e ‘idades’

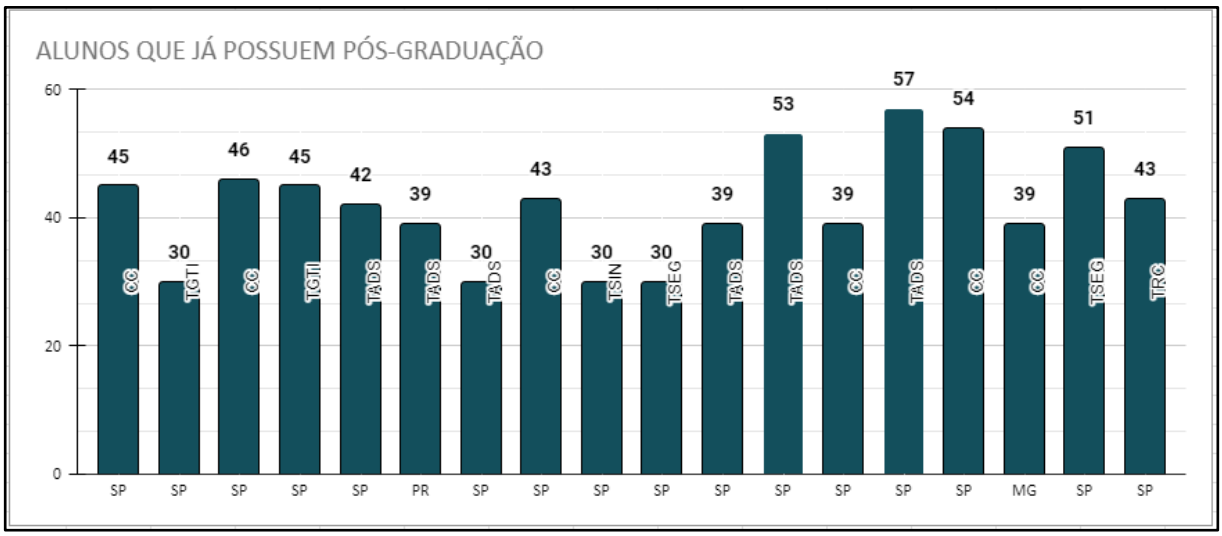


Fonte: Autora (2023)

Na população de alunos que já fizeram um curso de pós-graduação é de 75 alunos. Deste total é possível observar que 12 alunos possuem 27 anos, o que equivale a 16% deste grupo. Em segundo aparecem os alunos de 26 anos, 8 ao todo, o que representa 10,6% da população e em terceiro os alunos de 28 anos, que somam 7 alunos, correspondendo a 9,3% do total. Com estes resultados é possível notar que a população ingressante ao curso de Data Science possui uma massa de alunos que se concentra entre os jovens de até 30 anos.

Na Figura 64 são apresentados os resultados da correlação entre os atributos “região”, “curso de pós-graduação” e “idade” para alunos que já tem pós-graduação.

Figura 64: Correlação entre os atributos ‘curso de Pós-graduação’, ‘idade’, ‘Região’ e ‘curso de graduação’



Fonte: Autora (2023)

Os resultados apresentados na Figura 64 revelam a distribuição dos alunos que já possuem uma pós-graduação. A partir desses resultados, observa-se quais cursos de graduação esses alunos concluíram e de qual estado eles são originários. Isso nos proporciona conhecimento específico sobre a origem e características de cada aluno matriculado. Dentro desse grupo, é possível identificar que a maioria dos egressos optou pelos cursos de Tecnologia em Análise e Desenvolvimento de Sistemas (TADS) e Ciência da Computação (CC) em suas graduações. Essa constatação leva a entender que os alunos que escolhem uma especialização em Data Science buscam seguir ou manter uma carreira com foco específico na área de desenvolvimento de sistemas, dentro do campo principal da Informática.

Dentre os alunos matriculados nesse grupo, 66,67% (12 alunos) de um total de 18 já possuíam pós-graduação. Essa informação indica uma preferência clara por cursos de graduação relacionados à área de Informática, o que evidencia o interesse desses alunos em aprofundar seus conhecimentos e habilidades na área de Data Science. Sob uma perspectiva regional específica, São Paulo continua a se destacar em relação aos demais estados, com 16 alunos, o que representa 88,89% do total. Esses números sugerem que São Paulo atrai um perfil de indivíduos que estão em

busca de elevar seu nível de escolaridade. Além disso, temos uma aluna do Paraná (PR), do sexo feminino, com 39 anos, e um aluno de Minas Gerais (MG), do sexo masculino, também com 39 anos, cada um representando 5,56% do total de pós-graduados apresentados.

Na Figura 65 são apresentadas as correlações entre os atributos “média final das disciplinas gerenciais (MFG)”, “média final das disciplinas técnicas (MFT)” e “média final total das disciplinas (MF)” de alunos com pós-graduação.

Figura 65: Correlação entre os atributos “média final das disciplinas gerenciais (MFG)”, “média final das disciplinas técnicas (MFT)” e “média final total das disciplinas (MF)” de alunos com pós-graduação

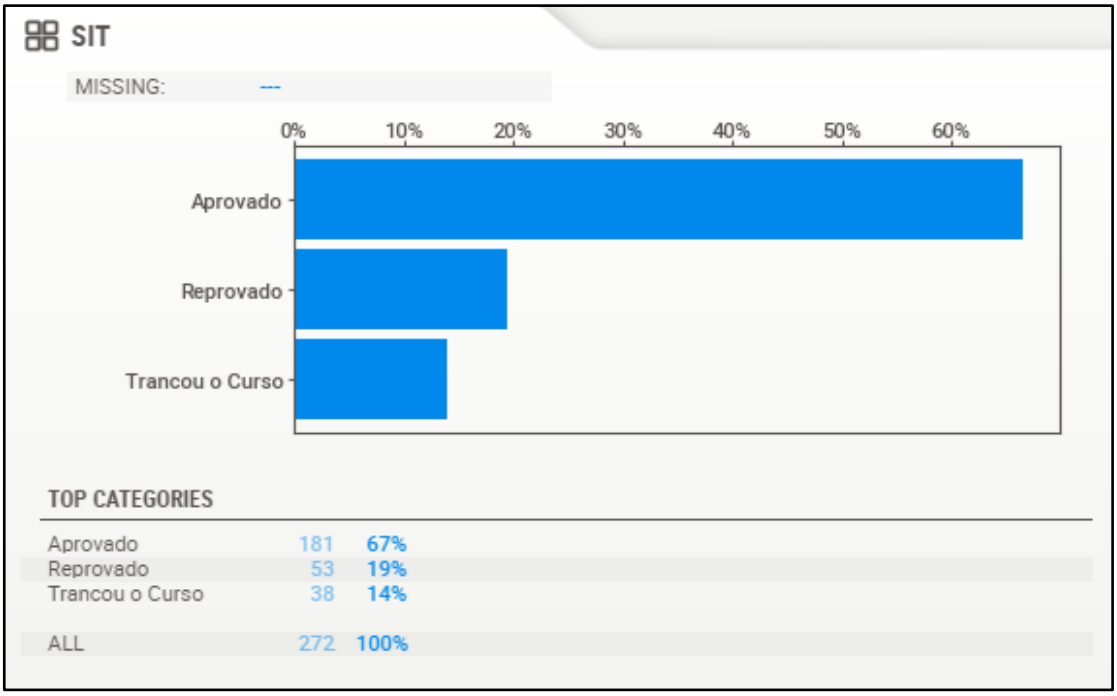
Fonte: Autora - *Software: Tableau* (2023)

Ao analisar o perfil dos alunos pós-graduados foi identificado um aluno em particular que teve um desempenho notável. Trata-se de um aluno do sexo masculino, com 54 anos, que concluiu o curso de Ciência da Computação (CC) em sua graduação e cursou o ensino médio na modalidade normal. No curso de Data Science, esse aluno obteve um desempenho excepcional nas disciplinas técnicas, alcançando uma média de 9,1. Nas disciplinas relacionadas a matérias gerenciais, obteve uma média de 5,5. Ao concluir o curso, o aluno conquistou uma média final de 8,1.

Em contrapartida, dois alunos não concluíram o curso, pois optaram por trancar seus estudos, ambos da Região de São Paulo e egressos do curso de Tecnologia em Análise e Desenvolvimento de Sistemas (TADS). Uma aluna de 42 anos e um aluno de 53 anos, que também cursaram o ensino médio na modalidade “normal”. Eles obtiveram um bom desempenho nas disciplinas técnicas, com média final técnica (MFT) respectivamente 7,1 e 8,8, porém nas disciplinas de assuntos gerenciais (MFG), a aluna obteve média 2,7 e o aluno não realizou as mesmas disciplinas. Resultando nas médias finais (MF) de 5,8 e 6,2 respectivamente.

Na Figura 66 é apresentado o resultado do atributo ‘situação’ dos alunos (aprovado, reprovado e trancou o curso).

Figura 66: Resultados de alunos no atributo “situação”

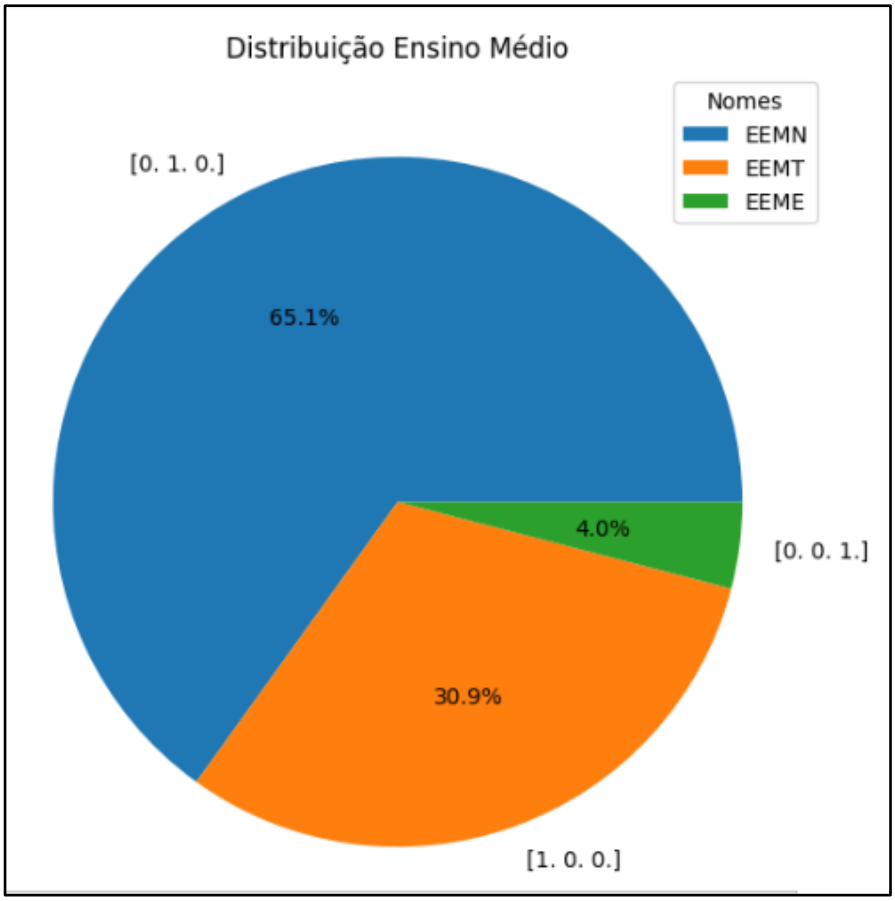


Fonte: Autora - Google Colab (2023)

O atributo "situação" representa o índice de aprovação do curso de pós-graduação. Observa-se que a maioria dos alunos obteve sucesso em seus estudos, sendo aprovados. Dos 272 alunos analisados, 181, representando aproximadamente 67% do total, alcançaram uma conclusão satisfatória. Por outro lado, 53 alunos, correspondendo a cerca de 19%, foram reprovados, o que indica um número significativo de reprovações no curso. Além disso, 38 alunos, ou aproximadamente 14%, optaram por interromper seus estudos e trancaram o curso.

Os alunos do curso de Data Science assim como os alunos de Governança em TI cursaram três tipos de ensino médio: EEMN (Escolaridade Ensino Médio Normal), EEMT (Escolaridade Ensino Médio Técnico) e EEME (Escolaridade Ensino Médio EJA). A Figura a seguir apresenta os resultados totais de alunos dos cursos de ensino médio (Normal, Técnico e EJA). A Figura 67 apresenta a distribuição dos alunos em relação ao tipo de ensino médio que cursaram.

Figura 67: Distribuição dos alunos Ensino Médio (EEMN, EEMT e EJA)



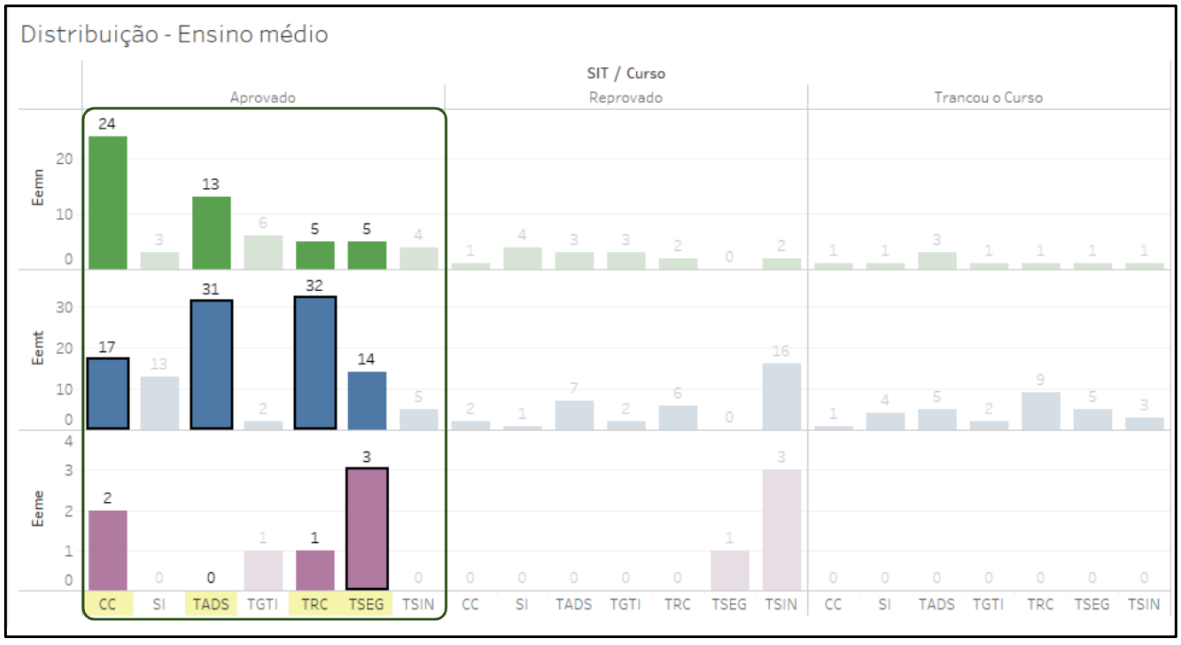
Fonte: Autora - Google Colab (2023)

A análise da distribuição dos dados revela que os alunos apresentam formações distintas no Ensino Médio. Aproximadamente 65,1% dos alunos cursaram o Ensino Médio Normal (EEMN), enquanto cerca de 30,9% cursaram o Ensino Médio Técnico (EEMT). Além disso, uma parcela menor, correspondente a 4% dos alunos, obteve sua formação no Ensino Médio por meio da modalidade de Educação de Jovens e Adultos (EEME).

Esses resultados refletem a diversidade de trajetórias educacionais dos alunos presentes no conjunto de dados, com uma proporção significativa tendo cursado o Ensino Médio Normal ou Técnico. A inclusão de alunos com formação em Educação de Jovens e Adultos também indica a presença de diferentes perfis de estudantes na análise. Essas informações são relevantes para obter percepções sobre o perfil educacional dos alunos e podem contribuir para a compreensão de possíveis correlações com outras variáveis presentes no conjunto de dados.

Na Figura 68 são apresentadas as correlações entre os atributos “curso”, “ensino médio” e “situação”.

Figura 68: Correlações entre os atributos “curso”, “ensino médio” e “situação”

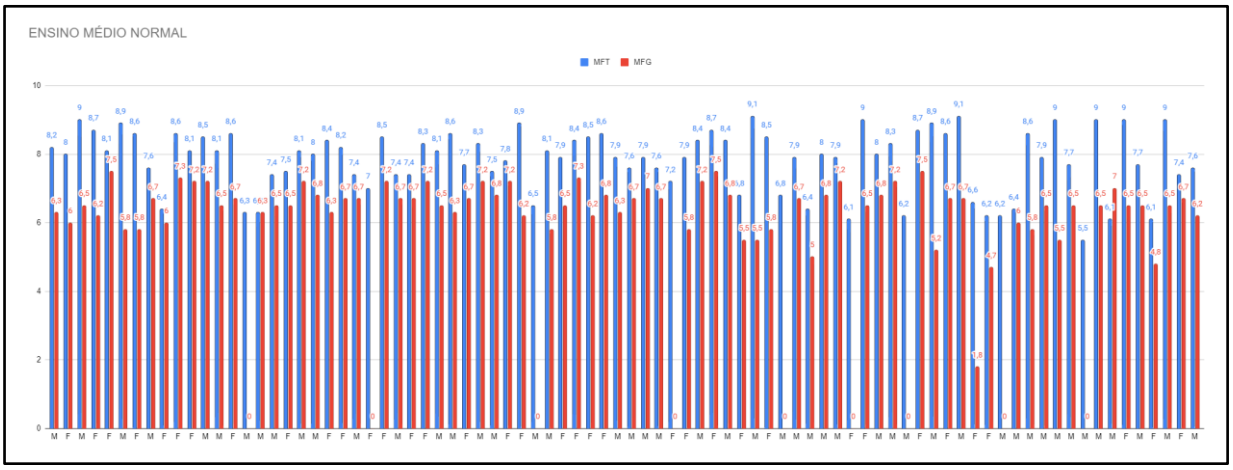


Fonte: Autora - *Software: Tableau* (2023)

É possível observar que os alunos provenientes do ensino médio normal estão concentrados principalmente no curso de Ciência da Computação (CC), com 24 alunos, equivale a 28,6% dos alunos desta modalidade, enquanto os alunos egressos do ensino médio técnico têm uma maior participação nos cursos de Tecnologia em Análise e Desenvolvimento de Sistemas (TADS) e Tecnologia em Redes de Computadores (TRC), com 31 e 32 alunos, respectivamente, um total de 63 dos 177 egressos desta modalidade, o que equivale a 35,6%. Já os alunos provenientes do ensino médio na modalidade EJA têm uma maior participação no curso de Tecnologia em Segurança da Informação (TSEG), com 3 alunos de um total de 11, que representa 27,3%, nesta modalidade EJA. Esses resultados indicam que os egressos do ensino médio técnico apresentam uma maior representatividade dentro do curso de Data Science.

Na Figura 69 é apresentado o desempenho dos alunos egressos de cada modalidade de ensino médio normal, separadamente por disciplina.

Figura 69: Análise exploratória - EEMN x Média



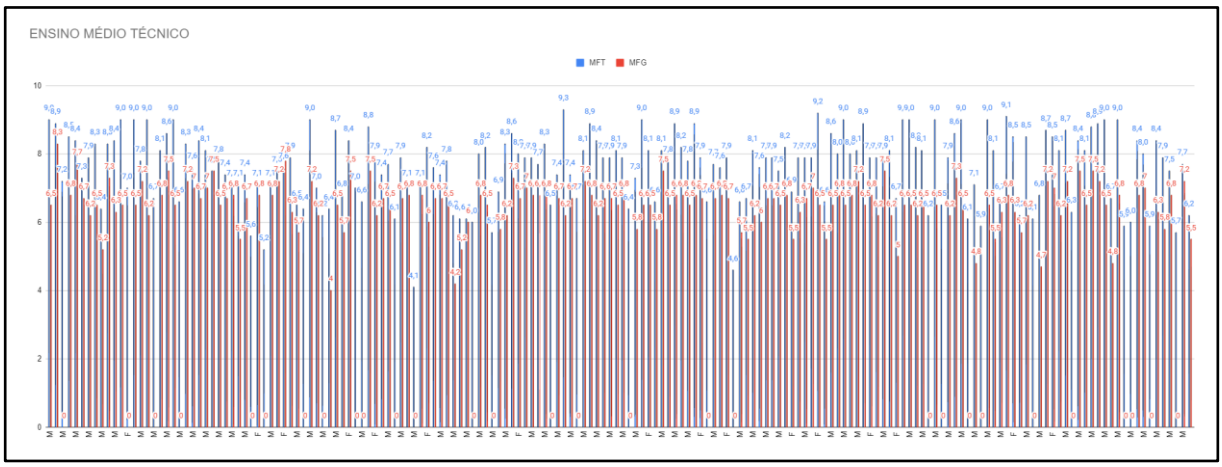
Fonte: Autora (2023)

Destaca-se a comparação entre as disciplinas gerenciais e técnicas na análise da performance dos alunos. As médias obtidas nas disciplinas técnicas merecem destaque, uma vez que, dos 84 alunos que cursaram essa modalidade, 68 deles (sendo 38 do sexo masculino e 30 do sexo feminino) alcançaram notas superiores a 7,0 nas disciplinas técnicas, enquanto obtiveram notas abaixo de 7,5 nas disciplinas gerenciais, o que corresponde a 81% dessa população. Vale ressaltar que a nota 7,5

foi a mais alta alcançada pelos alunos provenientes do ensino médio normal nas disciplinas gerenciais.

A Figura 70 apresenta as correlações entre os atributos “ensino médio técnico” e “média das disciplinas (técnica e gerencial)”.

Figura 70: Correlações entre os atributos “ensino médio técnico” e “média das disciplinas (técnica e gerencial)”



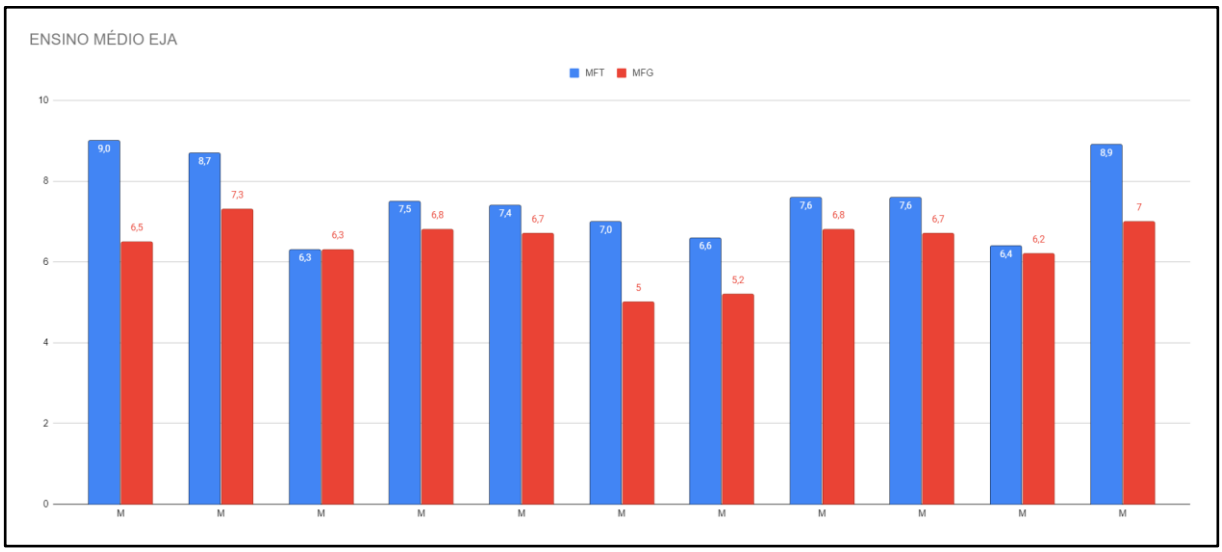
Fonte: Autora (2023)

Na Figura 70, são apresentados os resultados do desempenho dos alunos provenientes do ensino médio técnico, comparando as disciplinas técnicas e gerenciais. Neste grupo, observa-se que um total de 147 alunos (sendo 111 do sexo masculino e 36 do sexo feminino), de um universo de 177 alunos, obtiveram notas iguais ou superiores a 7,0 nas disciplinas técnicas, o que representa um desempenho superior à média de aprovação de 83,1% nessas disciplinas específicas relacionadas a desenvolvimento e assuntos técnicos.

Além disso, um total de 74 alunos (sendo 58 homens e 16 mulheres) obteve notas igual ou maior que 7,0 nas disciplinas gerenciais, representando 41,8% do total de estudantes analisados. Esses resultados indicam um desempenho notável tanto nas disciplinas técnicas quanto nas disciplinas gerenciais por parte dos alunos provenientes do ensino médio técnico.

A Figura 71 apresenta as correlações entre os atributos “ensino médio EJA” e “média das disciplinas (técnica e gerencial)”.

Figura 71: Correlações entre os atributos “ensino médio EJA” e “média das disciplinas (técnica e gerencial)”



Fonte: Autora (2023)

Os alunos provenientes do ensino médio Educação de Jovens e Adultos (EJA) compõem um grupo de menor tamanho, composto exclusivamente por estudantes do sexo masculino. Neste grupo, observa-se que os alunos obtiveram médias destacadas tanto nas disciplinas técnicas, com notas variando entre 6,3 e 9,0, quanto nas disciplinas gerenciais, com notas variando entre 5,0 e 7,3.

Esses resultados revelam um desempenho mais expressivo por parte dos alunos do ensino médio EJA nas disciplinas que são mais diretamente relacionadas a assuntos específicos e técnicos da área de estudo. Essa constatação sugere um maior domínio e habilidade dos alunos nessas matérias específicas, o que pode ser atribuído ao seu interesse e experiência prévia no campo de estudo abordado pelo curso.

Com base nos resultados obtidos para os alunos provenientes do ensino médio normal, ensino médio técnico e ensino médio EJA, é possível observar uma tendência consistente de desempenho superior nas sete disciplinas técnicas do curso de pós-graduação em Data Science. Essas disciplinas são caracterizadas por abordarem conteúdos mais técnicos e específicos da área de Data Science.

Essa constatação indica que os alunos apresentaram maior domínio e uma performance mais elevada nas disciplinas que envolvem conhecimentos técnicos e práticos relacionados à área de estudo. Isso sugere que os alunos,

independentemente do tipo de formação no ensino médio, estão mais aptos a lidar com os aspectos mais técnicos e especializados exigidos pelo curso de Data Science.

Esses resultados podem ser interpretados como uma indicação positiva do preparo e competência dos alunos em lidar com os desafios e demandas de um programa de pós-graduação em Data Science, especialmente nas áreas mais técnicas e específicas do campo. Isso reforça a importância de um embasamento sólido em conhecimentos prévios e habilidades técnicas para o sucesso acadêmico nesse domínio.

No que diz respeito ao desempenho dos egressos de cursos de graduação mencionados anteriormente, é perceptível que o curso de pós-graduação em questão apresenta um desempenho mais favorável para os alunos provenientes do curso de TADS (Tecnologia em Análise e Desenvolvimento de Sistemas). Esse grupo é composto por um total de 44 alunos, dos quais 30 são do gênero masculino e 14 do gênero feminino, representando uma taxa de aprovação de 16,2% no referido curso.

Ao analisar o atributo de gênero, é notável que os alunos do gênero masculino se destacam em relação ao curso de CC (Ciência da Computação), com 32 alunos aprovados, o que corresponde a 11,8% de sucesso no curso. Esses alunos possuem idades variando entre 25 e 72 anos, revelando a presença de um número significativo de indivíduos que concluíram sua graduação em Bacharelado e decidiram retornar aos estudos em uma fase mais avançada da vida ou que buscam aprimorar suas habilidades e conhecimentos em uma segunda especialização.

Além disso, o curso de TADS também registra um número expressivo de alunos aprovados, totalizando 30 indivíduos e representando 11% de aprovação. Esses alunos possuem idades entre 22 e 57 anos. Da mesma forma, o curso de TRC (Tecnologia em Redes de Computadores) conta com 29 alunos aprovados, o que equivale a uma taxa de aprovação de 10,7%. Os alunos desse curso possuem idades variando entre 22 e 59 anos. Esses dados revelam a presença de um público masculino mais experiente na maioria dos cursos analisados.

Essas informações evidenciam a diversidade de trajetórias educacionais e profissionais dos alunos envolvidos, bem como a busca por aprimoramento e especialização em uma área específica do conhecimento. Além disso, destacam a

importância de considerar fatores como idade, experiência e formação prévia ao analisar o desempenho dos alunos em um contexto de pós-graduação em Data Science.

No contexto da análise dos alunos que já possuíam uma pós-graduação, destacou-se um perfil de aluno que apresentou um desempenho excepcional. Trata-se de um aluno do sexo masculino, com 54 anos, que concluiu o curso de Ciência da Computação (CC) em sua graduação e cursou o ensino médio na modalidade normal. No curso de Data Science, esse aluno demonstrou um desempenho notável nas disciplinas técnicas, obtendo uma média de 9,1. Já nas disciplinas relacionadas a assuntos gerenciais, sua média foi de 5,5. Ao final do curso, ele conquistou uma média final de 8,1, refletindo sua habilidade e comprometimento.

Por outro lado, houve a ocorrência de dois alunos que não concluíram o curso, optando por interromper seus estudos. Ambos são provenientes da Região de São Paulo e egressos do curso de Tecnologia em Análise e Desenvolvimento de Sistemas (TADS). A aluna, com 42 anos, e o aluno, com 53 anos, também cursaram o ensino médio na modalidade "normal". Embora tenham apresentado um desempenho satisfatório nas disciplinas técnicas, com médias finais técnicas (MFT) de 7,1 e 8,8, respectivamente, nas disciplinas voltadas para assuntos gerenciais (MFG), a aluna obteve uma média de 2,7 e o aluno optou por não realizar tais disciplinas. Isso resultou em médias finais (MF) de 5,8 e 6,2, respectivamente.

Esses resultados ilustram a diversidade de trajetórias e desempenhos encontrados entre os alunos analisados no contexto da pós-graduação em Data Science. Enquanto um aluno demonstrou uma excelência notável em suas habilidades técnicas, outros dois alunos enfrentaram dificuldades que culminaram na interrupção de seus estudos. Essas situações ressaltam a importância de considerar fatores individuais e circunstanciais ao avaliar o desempenho dos alunos em um programa de pós-graduação.

Perfis a serem considerados

Como resultados preliminares gerais pôde-se evidenciar diferentes perfis de alunos como indicação de prognóstico de perfil de aluno ingressante no curso de pós-

graduação em Data Science, considerando-se os atributos escolaridade, gênero e idade em cruzamento com os resultados obtidos, conforme os dados da base de dados acadêmica processada.

Os resultados alcançados até o momento indicam um elevado índice de desempenho de alunos egressos de cursos de Graduação TADS (Tecnologia em Análise e Desenvolvimento de Sistemas) e Ensino Técnico, na faixa de até 35 anos. Também foi observada uma tendência consistente de desempenho elevado nas disciplinas iniciais do curso, que se referem as disciplinas de assuntos mais técnicos, demonstrando um bom domínio dos conteúdos específicos desta área. Esses resultados evidenciam a importância do embasamento sólido e do interesse prévio na área de estudo para o sucesso acadêmico em Data Science.

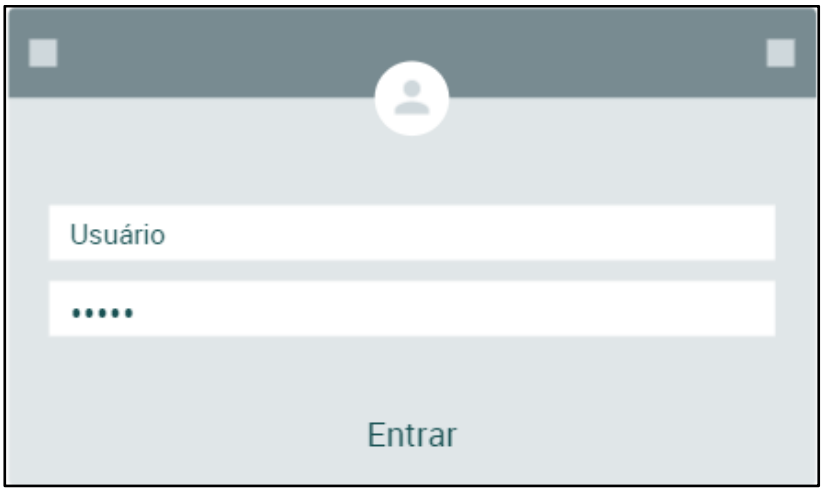
Observando os alunos pós-graduação, destacou-se o curso de graduação CC (Ciência da Computação) com alunos do sexo masculino, idade de média de 50 anos. Esses dados revelam a presença de um público masculino mais experiente na maioria dos cursos analisados. A partir deste perfil de maior aderência e melhor desempenho, pode-se proporcionar diretrizes para que os gestores da instituição de ensino superior tomem decisões mais acertadas para a preparação, atualização e divulgação dos cursos de ensino superior aos públicos-alvo mais pertinentes a cada curso específico.

5. SOLUÇÃO AUTOMATIZADA - PASSO A PASSO DO GESTOR PARA UTILIZAÇÃO DAS TELAS DA SOLUÇÃO DESENVOLVIDA

A solução automatizada elaborada nesta dissertação é uma ferramenta desenvolvida com o propósito de realizar análises de dados acadêmicos e formular um prognóstico do perfil de alunos ingressantes em cursos superiores, visando apoiar a gestão acadêmica. Seu objetivo é fornecer aos gestores uma visão geral das informações coletadas no banco de dados acadêmicos da instituição. O público-alvo dessa ferramenta são os gestores acadêmicos e administrativos das instituições de ensino superior.

Ao utilizar a solução automatizada a partir dos passos expostos a seguir, esses gestores poderão tomar decisões embasadas em informações consolidadas, tanto no que diz respeito à oferta de cursos e disciplinas, quanto à seleção de candidatos mais adequados a cada programa de ensino de pós-graduação. A solução proposta visa contribuir para a melhoria da eficiência da instituição, permitindo que ela ofereça uma educação mais alinhada às necessidades e expectativas dos alunos e do mercado. As imagens das telas foram desenvolvidas no software Fuid UI. Para início apresenta-se a tela de “*Login*”.

Figura 72: Solução - ‘Tela de Login’

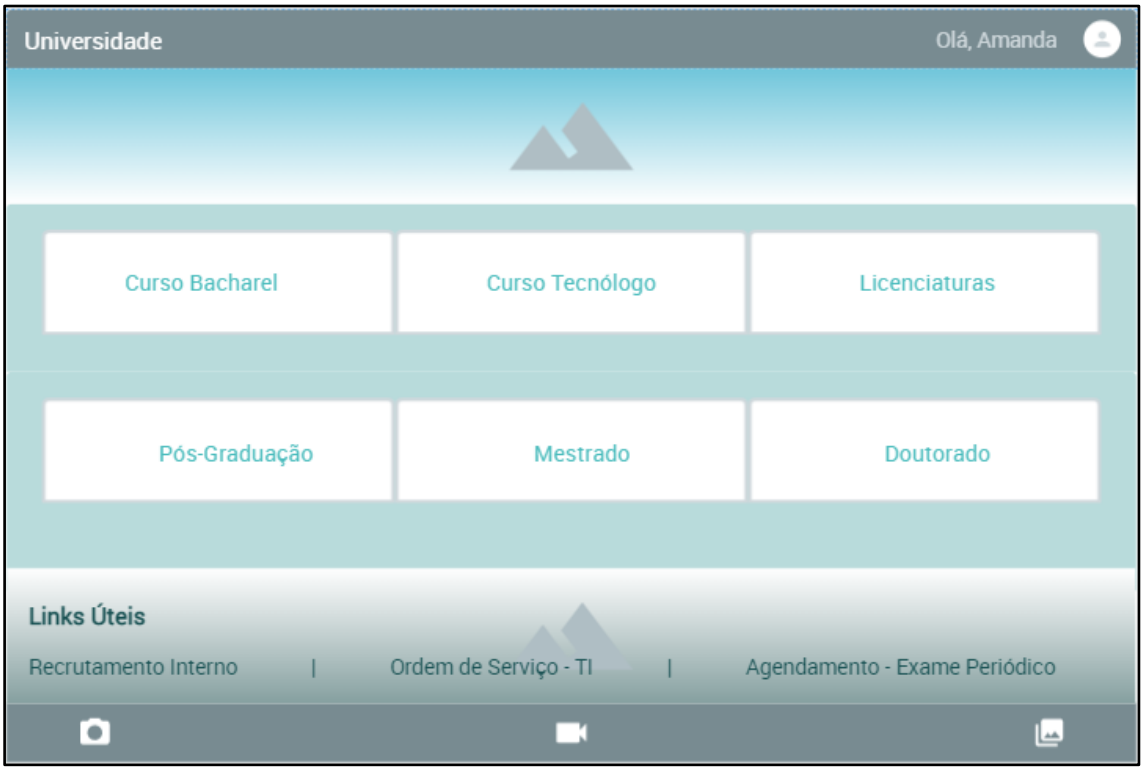


Fonte: Autora 2023

Após o usuário inserir seu nome de usuário e senha, ele será direcionado para a tela principal, onde estarão disponíveis as opções de cursos para consulta. Nessa

tela, o usuário poderá selecionar o curso desejado e acessar informações relevantes sobre ele.

Figura 73: Solução - ‘Tela de Consulta Principal’



Fonte: Autora 2023

Na tela principal, o usuário terá a opção de selecionar o tipo de curso a ser analisado pelo gestor ou atualizado pelo professor, levando em consideração as diferentes modalidades disponíveis, como presencial, semipresencial ou EAD (Educação a Distância). Cada modalidade apresentará suas respectivas opções de cursos para análise ou atualização.

Figura 74: Solução - ‘Tela de Consulta Bacharel (Modalidades)’

Fonte: Autora 2023

Ao selecionar uma modalidade específica, o usuário será direcionado para uma tela que exibirá os cursos disponíveis nessa modalidade. A partir daí o gestor poderá escolher um curso específico para análise, enquanto o professor poderá selecionar o curso no qual deseja fazer suas atualizações.

Figura 75: Solução - ‘Tela de Consulta Bacharel (Turmas)’

Fonte: Autora 2023

Ao selecionar um curso na tela principal, o usuário será direcionado para uma lista que exibirá todas as turmas disponíveis daquele curso na respectiva modalidade selecionada. Essa lista permitirá ao usuário visualizar e acessar informações específicas sobre cada turma do curso em questão.

Figura 76: Solução - ‘Tela de Consulta Bacharel (Funcionalidades)’

Fonte: Autora 2023

Dentro da opção "Turma", para uso do professor, estão localizadas as opções para atualização do desempenho.

- **“Atualizar dados da turma”:** onde o professor irá lançar as notas, faltas e fazer suas anotações;
- **"Planos de aula":** opção permite ao professor acessar os planos de aula do curso, fornecendo um cronograma detalhado das aulas, tópicos abordados, materiais de leitura e atividades relacionadas.
- **"Grade curricular":** opção permite ao professor acessar a grade curricular do curso, visualizando as disciplinas que compõem o currículo, suas descrições, carga horária e requisitos;
- **"Avaliações e trabalhos":** opção exibe as avaliações e trabalhos atribuídos aos alunos do curso, permitindo ao professor visualizar as datas de entrega, critérios de avaliação e registrar as notas correspondentes.
- **"Calendário acadêmico":** opção permite ao professor acessar o calendário acadêmico do curso, visualizando datas importantes, como períodos de matrícula, prazos de entrega de trabalhos, datas de provas e feriados.

- **"Estatísticas do curso":** opção exibe estatísticas relacionadas ao desempenho do curso, como média geral das notas, taxa de aprovação, distribuição de notas, entre outros indicadores relevantes.
- **"Recursos de ensino":** opção fornece acesso a recursos de ensino específicos do curso, como apresentações de slides, materiais didáticos, vídeos, listas de leitura recomendada, entre outros.
- **"Biblioteca virtual":** opção dá acesso à biblioteca virtual, onde o professor pode pesquisar e encontrar materiais acadêmicos relevantes para o curso, como artigos científicos, livros e periódicos.

Figura 77: Solução - 'Tela de Consulta Bacharel (Funcionalidades)'

Fonte: Autora 2023

Clicando em "**Atualizar dados da turma",** o professor irá ser direcionado à esta tela onde poderá fazer as atualizações necessárias ao perfil dos alunos, como: atribuir nota, registrar falta e atualizar o perfil do aluno, são fornecidos seguintes botões:

- **"Atribuir nota":** permite que o professor atribua uma nota a uma atividade, trabalho, prova ou qualquer outro componente de avaliação do aluno. Ao clicar

nesse botão, o professor pode inserir a nota correspondente e salvá-la no sistema.

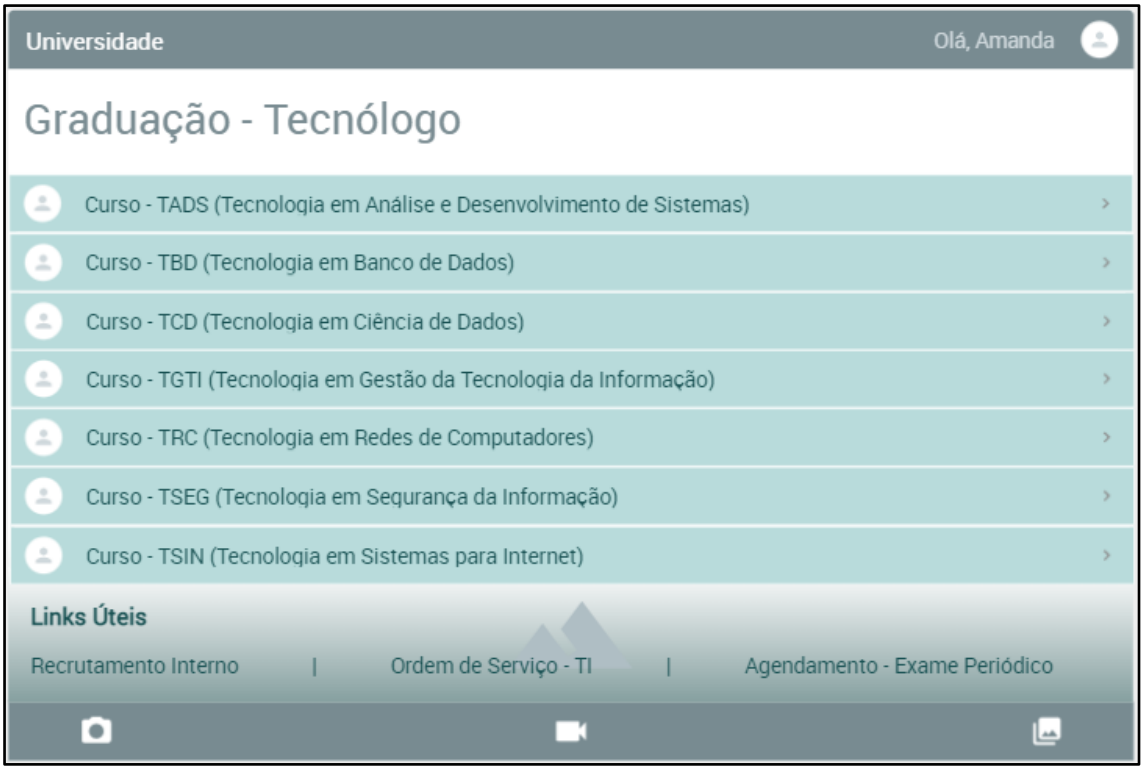
- **“Registrar falta”:** permite que o professor registre a falta de um aluno em uma aula, seminário, laboratório ou qualquer outra atividade acadêmica. Ao clicar nesse botão, o professor pode selecionar a data e a atividade específica em que o aluno faltou e salvar a informação.
- **“Atualizar perfil”:** permite que o professor atualize o perfil do aluno com informações relevantes, como observações, feedbacks, sugestões ou qualquer outra informação pertinente. Ao clicar nesse botão, o professor pode inserir as informações necessárias e salvar as alterações no perfil do aluno.
- **“Observações”:** permite que o professor faça uma anotação referente a turma, como uma atividade, ou tema que será trabalhado com a turma posteriormente, ou até um feedback adicional para salvar no sistema.
- **“Ver histórico de notas”:** permite que o professor acesse o histórico de notas do aluno, visualizando todas as notas atribuídas ao longo do período letivo. Isso pode ajudar o professor a ter uma visão geral do desempenho do aluno em diferentes atividades.
- **“Ver registro de faltas”:** permite que o professor acesse o registro de faltas do aluno, visualizando todas as faltas registradas ao longo do período letivo. Isso pode auxiliar o professor a identificar padrões de ausência e tomar as medidas necessárias.

O mesmo ocorre com os cursos de Tecnólogo, como mostram as próximas figuras:

Figura 78: Solução - ‘Tela de Consulta Tecnólogo (Modalidades)’

Fonte: Autora 2023

Figura 79: Solução - ‘Tela de Consulta Tecnólogo (Cursos)’



Fonte: Autora 2023

Figura 80: Solução - ‘Tela de Consulta Tecnólogo (Turmas)’

Fonte: Autora 2023

Figura 81: Solução - ‘Tela de Consulta Tecnólogo (Funcionalidades)’

Fonte: Autora 2023

Para fornecer ao gestor informações adicionais e permitir a análise do curso, turma ou aluno, está disponível a seção **"Estatísticas do curso"**. Nessa seção, é possível encontrar análises estatísticas com resultados que podem ser avaliados individualmente para cada aluno, em termos do desempenho total da turma ou em comparação com outras turmas do mesmo curso, bem como entre diferentes cursos.

Figura 82: Solução - 'Tela de Estatística do Curso'

Fonte: Autora 2023

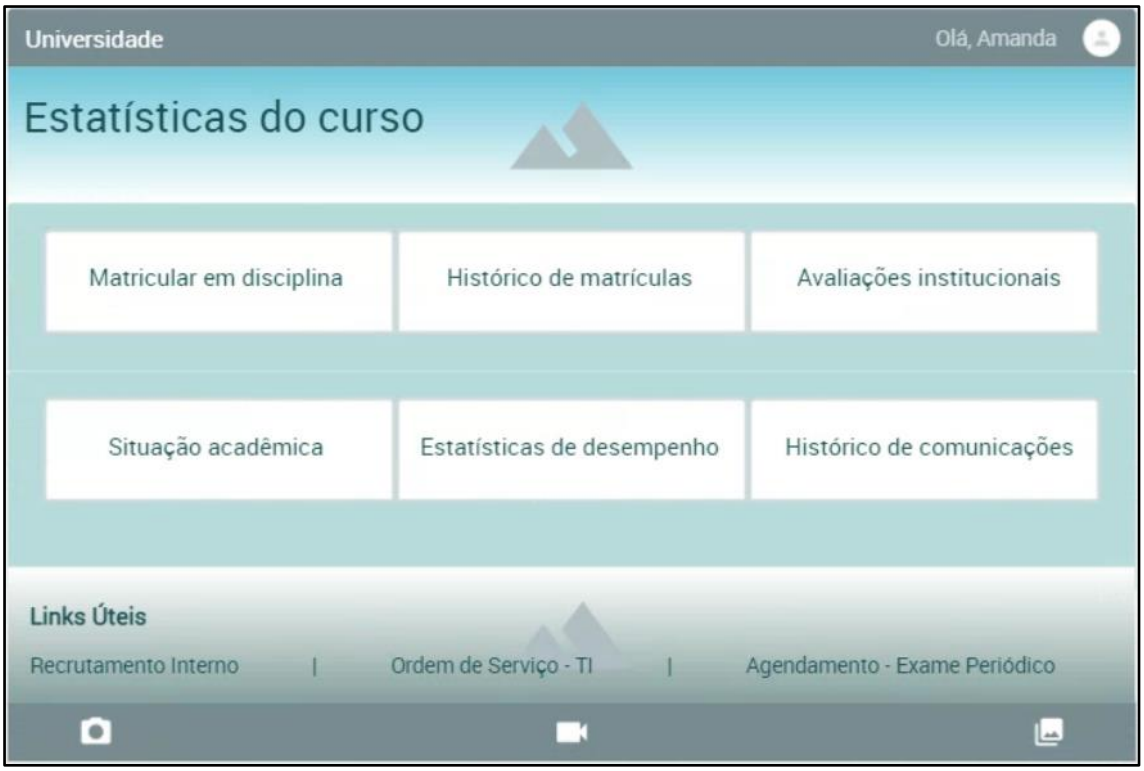
Estatísticas do curso possui as seguintes opções:

- **"Visualizar perfil completo":** permite que o gestor acesse todas as informações detalhadas e relevantes do perfil do aluno, incluindo informações pessoais, histórico acadêmico, cursos matriculados, notas, atividades extracurriculares, entre outros.
- **"Histórico acadêmico":** permite que o gestor acesse o histórico acadêmico do aluno, incluindo disciplinas cursadas, notas obtidas, média geral, carga horária cumprida, entre outros detalhes relevantes.
- **"Atividades extracurriculares":** permite que o gestor visualize as atividades extracurriculares em que o aluno participa ou participou, como clubes, equipes

esportivas, grupos de estudo, entre outros. Isso pode servir para avaliar o envolvimento do aluno na vida universitária.

- **"Informações pessoais":** fornece acesso às informações pessoais do aluno, como nome, data de nascimento, endereço, contato de emergência, entre outras informações necessárias para contato e identificação.
- **"Avaliações e feedbacks":** permite que o gestor acesse avaliações e feedbacks recebidos pelo aluno, tanto de professores como de outros alunos. Isso pode fornecer informações adicionais sobre o desempenho e a conduta do aluno.
- **"Progresso do curso":** exibe o progresso do aluno em relação ao curso que ele está matriculado. Pode incluir informações sobre disciplinas concluídas, disciplinas pendentes, créditos obtidos e requisitos para a graduação.

Figura 83: Solução - 'Tela de Estatística do Curso'



Fonte: Autora 2023

Ao rolar a página, você encontrará uma segunda tela que oferece uma variedade de opções adicionais de análise.

"Matricular em disciplina": permite que o gestor matricule o aluno em uma disciplina

específica, caso seja necessário fazer ajustes em sua grade curricular.

"Histórico de matrículas": exibe o histórico de matrículas do aluno, mostrando todas as disciplinas em que ele já esteve matriculado ao longo do tempo.

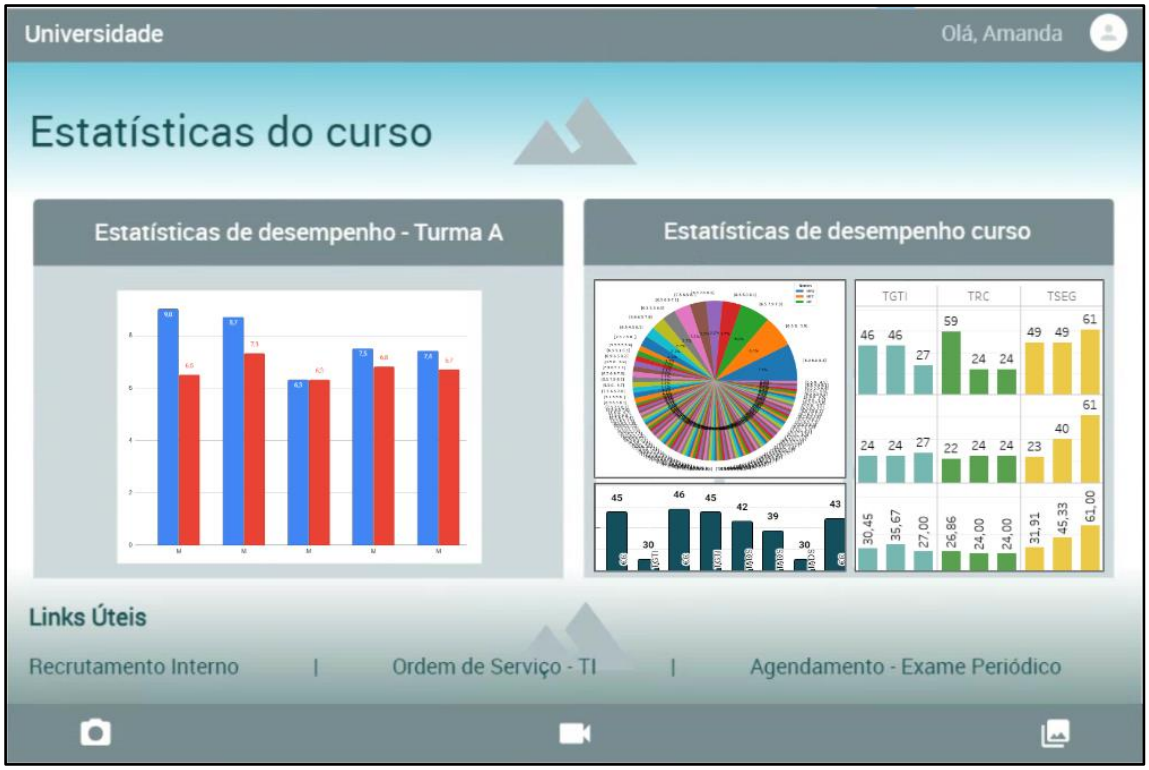
"Avaliações institucionais": permite que o gestor acesse os resultados das avaliações institucionais realizadas pelos alunos, como pesquisas de satisfação, feedbacks sobre a qualidade dos cursos, entre outros.

"Situação acadêmica": fornece uma visão geral da situação acadêmica do aluno, incluindo informações sobre status de matrícula, pendências, trancamentos, afastamentos, entre outros.

"Estatísticas de desempenho": apresenta estatísticas e métricas relacionadas ao desempenho do aluno, como média de notas, taxa de reprovação, taxa de evasão, entre outros indicadores relevantes.

"Histórico de comunicações": permite que o gestor acesse o histórico de comunicações realizadas com o aluno, como mensagens enviadas, notificações recebidas e interações anteriores.

Figura 84: Solução - ‘Tela de Resultados’



Fonte: Autora 2023

Figura 85: Solução - ‘Tela de Funcionalidades’

Fonte: Autora 2023

Esta tela oferece uma ampla variedade de escolhas de atributos para análise de correlações, permitindo que o gestor explore diferentes combinações de variáveis. Por exemplo, ao considerar as correlações entre gênero, curso de graduação e situação do aluno, o gestor pode identificar possíveis associações entre esses fatores, como diferenças de desempenho ou de taxa de aprovação entre gêneros em cursos específicos. Essa análise pode fornecer percepções sobre possíveis desigualdades de gênero no contexto acadêmico.

Além disso, o gestor pode investigar a relação entre gênero, curso de graduação e idade dos alunos. Essa análise permite examinar se existem diferenças de idade entre os gêneros em determinados cursos, o que pode indicar padrões de ingresso ou retenção de alunos em diferentes faixas etárias. Essas informações podem ser úteis para aprimorar estratégias de captação e engajamento de alunos, bem como para identificar possíveis barreiras ou oportunidades específicas para grupos demográficos. Seguem as descrições de cada atributo:

"Gênero x Curso de Graduação x Situação": Permite analisar a relação entre o gênero dos alunos, o curso de graduação que estão cursando e a situação acadêmica,

como aprovado, reprovado ou trancou o curso.

"Gênero x Curso de Graduação x Idade": Permite investigar a relação entre o gênero dos alunos, o curso de graduação e a faixa etária dos mesmos, podendo identificar possíveis padrões de idade entre os gêneros em cursos específicos.

"Pós-graduação x Idades": Analisa a relação entre alunos pós-graduados e suas faixas etárias, possibilitando identificar padrões de idade entre os estudantes que optaram por realizar mais um curso de pós-graduação.

"Curso de Pós-graduação x Idade x Região x Curso de graduação": Explora a relação entre alunos que já possuem pós-graduação, a faixa etária, a região geográfica em que estão localizados e o curso de graduação prévio. Essa análise permite compreender a diversidade de características dos alunos em diferentes programas de pós-graduação.

"EEMN x Médias das disciplinas": Relaciona a modalidade de ensino médio dos alunos (EEMN) com as médias obtidas nas disciplinas, permitindo verificar se existe alguma correlação entre o tipo de ensino médio e o desempenho acadêmico nas matérias.

"EEMT x Médias das disciplinas": Faz uma análise semelhante ao item anterior, mas considerando a modalidade de ensino médio técnico (EEMT) em relação às médias obtidas nas disciplinas.

"EEME x Médias das disciplinas": Relaciona a modalidade de ensino médio do tipo EEME com as médias obtidas nas disciplinas, permitindo avaliar se essa modalidade de ensino influencia o desempenho acadêmico nas matérias.

"Aprovados x Curso de Graduação": Explora a relação entre os alunos aprovados e o curso de graduação que eles estão cursando, possibilitando identificar quais cursos têm maior taxa de aprovação.

"Reprovados x Curso de Graduação": Essa análise investiga a relação entre os alunos reprovados e o curso de graduação que eles estão cursando. Permite identificar quais cursos têm uma maior incidência de reprovações, possibilitando a identificação de possíveis desafios ou dificuldades enfrentadas pelos estudantes em determinadas áreas de estudo.

"Trancou o curso x Curso de Graduação": Essa análise explora a relação entre os alunos que optaram por trancar o curso e o curso de graduação que eles estavam matriculados. Permite identificar quais cursos têm uma maior taxa de desistência ou

interrupção de estudos, fornecendo percepções sobre possíveis razões para a tomada dessa decisão, como falta de interesse, incompatibilidade com a área de estudo ou outras circunstâncias pessoais. Nas telas de resultados serão geradas visualizações gráficas correspondentes às análises selecionadas pelo gestor.

Figura 86: Solução - ‘Tela de Resultados’

Fonte: Autora 2023

Por fim, a solução automatizada proposta oferece ao gestor a flexibilidade de explorar uma ampla gama de atributos, permitindo uma compreensão mais aprofundada das correlações presentes nos dados acadêmicos. Essa variedade de escolhas oferece a oportunidade de identificar possíveis desigualdades de gênero, diferenças de idade entre os alunos, padrões de ingresso e retenção, e interações complexas entre diferentes variáveis. Compreender essas correlações e padrões pode auxiliar os gestores na tomada de decisões estratégicas, no desenvolvimento de políticas educacionais mais inclusivas e no direcionamento de ações específicas para cada segmento de alunos.

6. GUIA DE ORIENTAÇÃO PARA O DESENVOLVIMENTO E IMPLANTAÇÃO DA SOLUÇÃO AUTOMATIZADA

No Quadro 8 são apresentadas as etapas de implantação da solução automatizada desenvolvida, bem como as ferramentas, bases de dados, operações e ações necessários e uma breve descrição destas.

Quadro 8: Guia de orientação para implementação da solução automatizada

Etapas da implantação	Ferramentas, bases, operações, ações	Descrição
1 - Coleta de Dados Acadêmicos	Dados adicionados ao longo do semestre, tais como notas, créditos de trabalhos, inseridos gradualmente pelo professor.	Durante o semestre, foram selecionados e adicionados ao Banco de Dados para análise os dados adicionais, como notas, créditos de trabalhos, cursos prévios anteriores ao curso atual. Esses dados foram incluídos no banco de dados visando uma análise mais abrangente e detalhada.
2 - Pré-processamento de Dados	* Limpeza dos dados; * Codificação de variáveis categóricas; * Normalização ou padronização de dados, com objetivo colocar as variáveis em uma escala comum.	* A limpeza dos dados refere-se ao processo de identificar e lidar com dados ausentes, inconsistentes ou incorretos. Isso pode envolver a exclusão de registros com valores ausentes, o preenchimento de valores faltantes com estimativas adequadas (como a média ou a mediana) ou a remoção de outliers que podem prejudicar a análise. * A codificação de variáveis categóricas é necessária quando temos variáveis com valores não numéricos, como categorias ou rótulos. Algoritmos de aprendizado de máquina geralmente requerem variáveis numéricas como entrada, portanto, é preciso converter as variáveis categóricas em representações numéricas adequadas. Isso pode ser feito por meio de técnicas como: *".fit_transform",* "*.get_dummies*" onde cada categoria é representada por uma nova variável binária, ou por meio de codificação ordinal, onde as categorias são mapeadas para valores numéricos ordenados. * A normalização ou padronização de dados é aplicada quando as variáveis possuem escalas diferentes e é necessário colocá-las em uma escala comum. A normalização ajusta os valores das variáveis para estarem em uma faixa específica, como 0 a 1, facilitando a comparação entre diferentes variáveis. A padronização, por sua vez, transforma as variáveis para terem média zero e desvio padrão igual a um, o que é útil em algoritmos sensíveis à escala das variáveis, como muitos algoritmos de aprendizado de máquina.
3 - Solução	* Aplicação do algoritmo de clusterização, como K-means, Gaussian Mixture Models ou Hierarchical clustering (Agrupamento Hierárquico);	Após o pré-processamento dos dados, a etapa seguinte é a aplicação de um algoritmo de clusterização, como K-means, Gaussian Mixture Models (GMM) ou Hierarchical clustering (Agrupamento Hierárquico). Esses algoritmos são utilizados para agrupar os dados em conjuntos ou clusters com base em suas características similares. * O algoritmo K-means é um método de clusterização amplamente utilizado que busca particionar os dados em K clusters, onde K é um valor pré-definido. Ele atribui cada ponto de dados ao cluster mais próximo com base nas

		distâncias entre os pontos e os centróides dos clusters. O objetivo é minimizar a soma dos quadrados das distâncias entre os pontos e seus centróides. * Os Gaussian Mixture Models (GMM) são algoritmos que assumem que os dados em cada cluster seguem uma distribuição Gaussiana (normal). Eles modelam a distribuição de probabilidade dos dados em cada cluster e, em seguida, atribuem pontos de dados ao cluster mais provável com base nas probabilidades calculadas. * O Hierarchical clustering (Agrupamento Hierárquico) é um algoritmo que constrói uma estrutura hierárquica de clusters. Ele pode ser dividido em dois tipos: aglomerativo e divisivo. O aglomerativo começa com cada ponto de dados como um cluster separado e, em seguida, combina os clusters mais próximos sucessivamente até formar um único cluster. O divisivo começa com todos os pontos como um único cluster e, em seguida, divide-os em subclusters sucessivamente. A escolha do algoritmo de clusterização depende das características dos dados e dos objetivos da análise. O K-means é rápido e eficiente para grandes conjuntos de dados, mas requer a definição prévia do número de clusters. Os GMMs são mais flexíveis e podem modelar distribuições mais complexas, mas podem ser computacionalmente mais intensivos. O Hierarchical clustering oferece uma visão hierárquica dos clusters, mas pode ser computacionalmente mais exigente em grandes conjuntos de dados. A aplicação desses algoritmos de clusterização permite agrupar os dados em conjuntos significativos com base em suas similaridades, ajudando na identificação de padrões e na compreensão da estrutura subjacente nos dados. Essa informação pode ser útil para tomada de decisões, segmentação de clientes, análise de mercado, entre outras aplicações.
	* Treinamento do modelo;	* O treinamento do modelo, envolve fornecer os dados de entrada ao algoritmo e permitir que ele aprenda a identificar os padrões e agrupamentos nos dados. Durante o treinamento, o modelo ajusta seus parâmetros com base nos dados fornecidos, a fim de otimizar sua capacidade de agrupar os dados de forma adequada.
	* Ajuste os parâmetros do algoritmo;	* O ajuste dos parâmetros do algoritmo é uma etapa importante no processo de treinamento do modelo. Cada algoritmo de clusterização tem seus próprios parâmetros que podem ser ajustados para melhor se adequar aos dados e aos objetivos da análise. Por exemplo, no algoritmo K-means, é necessário definir o número de clusters (K), enquanto nos Gaussian Mixture Models (GMM), é preciso ajustar o número de componentes Gaussianos.
	* A representação dos diferentes perfis de alunos identificados pelo modelo;	* Após o treinamento e ajuste dos parâmetros, o modelo é capaz de representar os diferentes perfis de alunos identificados. Cada cluster formado pelo modelo representa um grupo de alunos com características semelhantes. Essas características podem ser atributos demográficos, desempenho acadêmico, envolvimento em atividades extracurriculares, entre outros. A representação dos perfis de alunos pode ser visualizada por meio de gráficos, tabelas ou outras formas de visualização, fornecendo uma compreensão clara dos grupos identificados pelo modelo.

	* Avaliação do modelo com uso de métricas como: - Coeficiente de Silhueta (Silhouette Coefficient), - Índice de Rand ajustado (Adjusted Rand Index - ARI), - Índice de Pureza (Purity Score) e Inércia; * Predição do perfil do aluno.	A representação dos diferentes perfis de alunos é uma etapa crucial, pois fornece percepções valiosas para a tomada de decisões e ações estratégicas. Com base nos perfis identificados, os gestores acadêmicos podem adaptar as estratégias de ensino, oferecer suporte acadêmico específico para cada grupo, desenvolver programas de retenção e aprimorar a experiência educacional dos alunos. Essa compreensão dos diferentes perfis de alunos também pode auxiliar na divulgação do curso, direcionando ações de marketing e comunicação para grupos específicos. As métricas específicas utilizadas nesta solução, fornecem uma medida objetiva da eficácia do modelo: * Coeficiente de Silhueta: O coeficiente de silhueta é uma medida que avalia a coesão e separação dos clusters. Ele varia de -1 a 1, onde valores mais próximos de 1 indicam que os pontos dentro de um cluster estão bem próximos uns dos outros e distantes dos pontos de outros clusters. * Índice de Rand ajustado: O índice de Rand ajustado é uma métrica que mede a similaridade entre os clusters obtidos pelo modelo e as classes verdadeiras dos dados. Ele varia de -1 a 1, onde valores mais próximos de 1 indicam maior concordância entre os clusters e as classes verdadeiras. * Índice de Pureza: O índice de pureza mede a pureza dos clusters em relação às classes verdadeiras. Ele calcula a proporção dos pontos do cluster que pertencem à classe mais frequente no cluster. O índice de pureza varia de 0 a 1, onde valores mais próximos de 1 indicam clusters mais puros. * Inércia: A inércia é uma medida específica para o algoritmo K-means. Ela representa a soma das distâncias quadráticas dos pontos para o *centróide* do cluster ao qual eles pertencem. O objetivo é minimizar a inércia, ou seja, ter clusters compactos e bem definidos. * A predição do perfil do aluno é útil para diversas finalidades, como personalizar o suporte acadêmico, adaptar o currículo e as estratégias de ensino, e fornecer orientações e recomendações individualizadas aos alunos. Com base no perfil previsto, os gestores acadêmicos podem tomar medidas adequadas para melhorar a experiência dos alunos e promover seu sucesso acadêmico.
4 - Apresentação dos Resultados	* Análises; * Resultados;	* Análises: Nesta etapa, os resultados obtidos são analisados em detalhes. Isso inclui examinar as características dos diferentes clusters identificados, entender os padrões e tendências encontrados nos dados e realizar comparações entre os grupos. É importante explorar as informações extraídas dos dados para obter percepções sobre os alunos e seu desempenho acadêmico. * Resultados: Os resultados da análise dos dados e da clusterização são apresentados de forma clara e objetiva. Isso pode ser feito por meio de tabelas, gráficos, visualizações ou outras representações visuais que facilitem a compreensão dos resultados. É importante destacar as principais descobertas e visões obtidas, destacando as diferenças e semelhanças entre os clusters e as implicações para a gestão acadêmica.

	* Recomendações.	* Recomendações: Com base nos resultados obtidos, são feitas recomendações relevantes para a gestão acadêmica. Isso pode incluir sugestões de ações a serem tomadas para melhorar o desempenho dos alunos, adaptar o currículo, oferecer suporte acadêmico personalizado, promover a estadia dos alunos, entre outras medidas. As recomendações devem ser baseadas nas percepções extraídas dos dados e nos padrões identificados nos clusters, visando aprimorar a experiência dos alunos e dar suporte a gestão acadêmica.

Fonte: Autora (2023)

7. CONCLUSÃO

A mineração de dados educacionais desempenha um papel preponderante na gestão acadêmica, possibilitando às instituições de ensino coletar e analisar dados com o intuito de obter percepções valiosas sobre o desempenho dos alunos, identificar padrões de comportamento e embasar decisões estratégicas. Por meio da análise dos dados educacionais é possível aprofundar a compreensão do perfil dos alunos, identificar suas necessidades e dificuldades, bem como avaliar o desempenho dos cursos e programas ofertados pela instituição.

A análise de dados educacionais possibilita a identificação de tendências e padrões de desempenho dos alunos, tais como taxas de aprovação, reprovação e evasão. Isso permite que os gestores acadêmicos identifiquem áreas problemáticas e desenvolvam estratégias visando aprimorar a qualidade do ensino, elevar a permanência dos alunos e promover maior satisfação entre os estudantes. Adicionalmente, a mineração de dados educacionais pode revelar fatores que influenciam o desempenho acadêmico.

Um aspecto de grande importância na mineração de dados educacionais é a capacidade de realizar prognósticos do perfil do aluno. Por meio da utilização dos dados armazenados no banco de dados da universidade é possível desenvolver modelos preditivos que auxiliam no prognóstico do desempenho futuro dos estudantes. Esses prognósticos fornecem aos gestores e às instituições de ensino subsídios valiosos para embasar a divulgação dos cursos, promover adaptações curriculares e desenvolver estratégias de suporte acadêmico. Ao compreender os fatores que influenciam o sucesso do aluno, as instituições podem implementar medidas preventivas e personalizadas, visando garantir a eficácia do processo de ensino-aprendizagem.

Com base nesse contexto, esta pesquisa propôs a utilização de uma solução computacional de mineração de dados educacionais para o desenvolvimento e validação de uma solução automatizada que visa fornecer suporte à gestão acadêmica na formulação de prognósticos do perfil de alunos ingressantes em cursos superiores. O objetivo da solução automatizada desenvolvida é utilizar técnicas de mineração de dados para analisar informações relevantes sobre os alunos, a fim de identificar padrões e tendências que possam auxiliar na compreensão de

características e necessidades dos estudantes desde o início de sua trajetória acadêmica. Essa solução automatizada pode contribuir significativamente para uma gestão mais eficiente e embasada em dados, fornecendo informações valiosas para a tomada de decisões estratégicas no contexto acadêmico.

A solução automatizada desenvolvida nesta dissertação é uma ferramenta abrangente para a análise de dados acadêmicos e previsão do perfil de alunos ingressantes em cursos superiores, com o objetivo de apoiar a gestão acadêmica. Cabe ressaltar que a solução automatizada desenvolvida se apoia em ferramentas gratuitas ou de baixo custo, o que o torna acessível a qualquer instituição que deseje implementá-lo.

Voltado para gestores acadêmicos e administrativos de instituições de ensino superior, a solução automatizada oferece uma estrutura que permite aos usuários selecionar o tipo de curso e modalidade desejados, bem como acessar informações específicas sobre cursos e turmas. Em especial, o segmento "Estatísticas do curso" da solução automatizada oferece opções como visualização do perfil completo do aluno, histórico acadêmico, atividades extracurriculares, informações pessoais, avaliações e feedbacks, progresso do curso, avaliações institucionais, situação acadêmica, estatísticas de desempenho entre outras opções, para que a gestão acadêmica possa tomar decisões embasadas e obter uma visão abrangente do desempenho dos alunos e das necessidades da instituição.

Os resultados verificados para o curso Governança em TI identificaram que alunos que concluíram o Ensino Médio Normal (EEMN) e cursaram na graduação o curso de Tecnologia em Gestão da Tecnologia da Informação (TGTI) apresentaram um desempenho acadêmico superior em relação a outros perfis de alunos. Esses alunos, em média, possuíam uma idade de 30 anos, eram do gênero feminino e já haviam realizado um curso de especialização prévio. Por outro lado, verificou-se que alunos provenientes do Ensino Médio Técnico (EEMT), que cursaram a graduação em Tecnologia em Sistemas para Internet (TSIN), com uma idade abaixo de 29 anos e do gênero masculino apresentaram um desempenho acadêmico inferior quando comparados a outros perfis. Além disso, constatou-se que esses alunos com menor desempenho não possuíam formação prévia em pós-graduação.

Quanto ao curso de Data Science, os resultados obtidos por meio da análise preliminar dos dados acadêmicos processados revelaram diferentes perfis de alunos com indicadores promissores para a previsão do perfil de futuros alunos ingressantes. Foi observado um alto desempenho por parte dos alunos egressos de curso de graduação em TADS (Tecnologia em Análise e Desenvolvimento de Sistemas) e provenientes do Ensino Médio Técnico, com idade até 35 anos. Além disso, os resultados apontaram uma tendência consistente de desempenho elevado nas disciplinas iniciais do curso, que abordam assuntos mais técnicos, indicando um bom domínio dos conteúdos específicos da área. A presença de alunos do sexo masculino, com idade média de 50 anos, provenientes do curso de CC (Ciência da Computação) e já pós-graduados, destaca o perfil de um público mais experiente e indica a necessidade de direcionar estratégias de preparação, atualização e divulgação dos cursos de ensino superior para públicos-alvo mais pertinentes. Esses resultados fornecem diretrizes valiosas para os gestores das instituições de ensino superior na tomada de decisões mais informadas e eficazes em relação aos cursos oferecidos, visando atrair os estudantes mais adequados e promover um ensino de qualidade.

Outro ‘produto’ importante da solução delineada nesta dissertação é o guia de orientação para implantação da solução automatizada, que descreve os passos para o desenvolvimento e implantação da solução elaborada. O guia apresenta as etapas a serem percorridas (1 - Coleta de dados acadêmicos; 2 - Pré-processamento de dados; 3 - Solução e 4 - Apresentação dos resultados), bem como as ferramentas, bases de dados, operações e ações necessários para tanto, além de uma breve descrição destas.

A solução automatizada desenvolvida está embasada em algoritmos de clusterização voltados a identificar e agrupar os alunos em conjuntos (*clusters*) com base em suas características (atributos) similares. Isso permite uma compreensão mais aprofundada dos perfis de alunos, facilitando assim a identificação de padrões e tendências nos dados acadêmicos disponíveis. A importância dessa análise de clusterização está em fornecer informações valiosas para a gestão acadêmica. Com base nos perfis identificados é possível personalizar o suporte acadêmico, adaptar o currículo e as estratégias de ensino, desenvolver programas de retenção e aprimorar a experiência educacional dos alunos. Além disso, as recomendações resultantes da

análise dos *clusters* podem direcionar ações estratégicas, como a segmentação de clientes e a análise de mercado.

A utilização dos algoritmos de clusterização empregados na solução automatizada desta dissertação permite explorar grandes volumes de dados de forma eficiente, identificando grupos de alunos com características semelhantes que podem não ser óbvias em uma análise manual. Isso proporciona uma compreensão mais profunda do corpo discente, permitindo a tomada de decisões embasadas em dados e ações mais direcionadas. Portanto, o objetivo da implantação da solução automatizada e a importância da análise de clusterização voltam-se a promover uma gestão acadêmica mais eficiente, oferecer suporte personalizado aos alunos, melhorar o desempenho acadêmico e a experiência educacional, além de contribuir para o planejamento estratégico e a tomada de decisões embasadas em dados.

Esta pesquisa apresenta contribuições para a Academia, pois expõe as possibilidades de aplicação de técnicas e ferramentas de mineração de dados educacionais para apoiar a gestão acadêmica. Por meio da análise dos dados, a solução automatizada desenvolvida permite identificar padrões e tendências no desempenho dos alunos, compreender suas características e necessidades, bem como realizar prognósticos do perfil dos estudantes ingressantes em cursos superiores.

A pesquisa apresenta contribuições também para as Instituições de Ensino Superior e seus gestores, fornecendo uma ferramenta de mineração de dados educacionais que permite uma gestão acadêmica mais eficiente e embasada em dados. Por meio da análise dos dados acadêmicos, a solução automatizada ora concebida possibilita a identificação de padrões e tendências no desempenho dos alunos, permitindo aos gestores compreender as necessidades e características dos estudantes. Isso facilita a personalização do suporte acadêmico, a adaptação curricular, o desenvolvimento de estratégias de retenção e aprimoramento da experiência educacional. Além disso, a ferramenta oferece uma visão abrangente do desempenho dos alunos e das necessidades da instituição, permitindo uma tomada de decisões mais embasada e estratégica. Com essas contribuições, as instituições de ensino podem melhorar a qualidade do ensino, aumentar a satisfação dos alunos e alcançar resultados mais efetivos em sua missão educacional. Os gestores também

se beneficiam ao terem acesso a informações valiosas para embasar suas decisões e promover uma gestão mais eficiente e direcionada.

Esta pesquisa apresenta algumas limitações a serem consideradas. Primeiramente, o estudo utilizou uma base de dados com um volume relativamente baixo de dados, o que pode limitar a representatividade dos resultados, não obstante sua fidedignidade. Além disso, as técnicas e métodos de mineração de dados educacionais utilizados para a clusterização dos alunos podem ter suas próprias limitações, como a escolha dos algoritmos utilizados e das variáveis consideradas na análise. Outra possível limitação é a dependência da qualidade e disponibilidade dos dados utilizados, incluindo a precisão e completude das informações coletadas. Além disso, é importante destacar que as conclusões e recomendações obtidas a partir da análise de clusterização são baseadas em padrões identificados nos dados, mas não necessariamente indicam relações causais.

Com base nesta pesquisa, algumas sugestões para pesquisas futuras podem ser indicadas. Primeiramente, seria interessante expandir o escopo da análise de clusterização para incluir um maior volume de dados e uma maior diversidade de variáveis, a fim de obter uma compreensão mais abrangente dos perfis de alunos. Além disso, poderiam ser exploradas diferentes técnicas de mineração de dados voltadas à clusterização, de modo a possibilitar a comparação dos resultados obtidos com os algoritmos utilizados neste estudo. Outra possibilidade seria investigar a aplicação de outras técnicas de mineração de dados voltadas à análise de associação ou classificação, para assim identificar outros padrões e relações entre os dados educacionais. Além disso, seria relevante avaliar o impacto das recomendações e ações derivadas da análise de clusterização na gestão acadêmica e no desempenho dos alunos, por meio de estudos de caso ou experimentos controlados. Por fim, seria válido explorar a utilização de abordagens de aprendizado de máquina e inteligência artificial para aprimorar a precisão das previsões e análises realizadas.

REFERÊNCIAS

ABU TAIR, Mohammed M.; EL-HALEES, Alaa M. Mining educational data to improve students' performance: a case study. **International Journal of Information**, v. 2, n. 2, 2012.

ADEKITAN, Aderibigbe Israel; SALAU, Odunayo. The impact of engineering students' performance in the first three years on their graduation result using educational data mining. **Heliyon**, v. 5, n. 2, p. e01250, 2019.

ALDOWAH, Hanan; AL-SAMARRAIE, Hosam; FAUZY, Wan Mohamad. Educational data mining and learning analytics for 21st century higher education: A review and synthesis. **Telematics and Informatics**, v. 37, p. 13-49, 2019.

ALEXANDER, Bryan et al. EDUCAUSE Horizon Report. **Higher Education Edition). Louisville, Co: Educause**, 2019.

ALQASEMI, Fahd, *et al*., Education Data Mining For Yemen Regions Based On Hierarchical Clustering Analysis. In: **2021 International Conference of Technology, Science and Administration (ICTSA)**. IEEE, 2021. p. 1-4.

ALSUWAIKET, Mohammed; BLASI, Anas H.; AL-MSIE'DEEN, Ra'Fat. Formulating module assessment for improved academic performance predictability in higher education. **arXiv preprint arXiv:2008.13255**, 2020.

ALYAHYAN, Eyman & DUSTEGOR, Dilek. Predicting Academic Success in Higher Education Literature Review and Best Practices. **International Journal of Educational Technology in Higher Education**. 17. 10.1186/s41239-020-0177-7, 2020

ANOOPKUMAR, M.; RAHMAN, AMJ Md Zubair. A Review on Data Mining techniques and factors used in Educational Data Mining to predict student amelioration. In: **2016 International Conference on Data Mining and Advanced Computing (SAPIENCE)**. IEEE, 2016. p. 122-133.

BACURAU, Rodrigo M.; LEAL, Brauliro G.; RAMOS, Ricardo A. Uma Abordagem para a Construção de Diagramas da UML Concomitante à Prototipação de Interface. 2022.

Barreto, Lucas; Monteiro, Edwin; Leitão, Gabriel; Bentes, Thays; Barreto, Raimundo. (2020). **Mineração de Dados Educacionais a partir da Interação de Alunos com uma Plataforma Educacional**. 10.5753/cbie.sbie.2020.1052.

BAKER, Phillip et al. Global trends and patterns of commercial milk-based formula sales: is an unprecedented infant and young child feeding transition underway?. **Public health nutrition**, v. 19, n. 14, p. 2540-2550, 2016.

BAKHSHINATEGH, Behdad, *et al.*, Educational data mining applications and tasks: A survey of the last 10 years. **Education and Information Technologies**, v. 23, n. 1, p. 537-553, 2018.

BECKER, Samantha Adams et al. **Horizon report 2018 higher education edition brought to you by educause**. EDUCAUSE, 2018.

BETTIN, Giovanna C.; GERALDI, Ricardo Theis; OLIVEIRAJR, Edson. Experimental Evaluation of the SMartyCheck Techinique for Inspecting Defects in UML Component Diagrams. **In: Simpósio Brasileiro De Qualidade De Software (SBQS)**, 17., 2018, Curitiba.

BRIGGS, Christopher; FAN, Zhong; ANDRAS, Peter. Federated learning with hierarchical clustering of local updates to improve training on non-IID data. In: 2020 **International Joint Conference on Neural Networks (IJCNN).** IEEE, 2020. p. 1-9.

BRITO, Daniel Miranda et al. Identificação de estudantes do primeiro semestre com risco de evasão através de técnicas de Data Mining. **Nuevas Ideas en Informática Educativa TISE**, p. 459-463, 2015.

BROCH, Caroline; BRESCHILIARE, Fabiane Castilho Teixeira; BARBOSA-RINALDI, Ieda Parra. A expansão da educação superior no Brasil: notas sobre os desafios do trabalho docente. **Avaliação: Revista da Avaliação da Educação Superior (Campinas)**, v. 25, p. 257-274, 2020.

BROOK, Cheryl; PEDLER, Mike. Action learning in academic management education: A state of the field review. **The International Journal of Management Education**, v. 18, n. 3, p. 100415, 2020.

CALVET LIÑÁN, Laura; JUAN PÉREZ, Ángel Alejandro. Educational Data Mining and Learning Analytics: differences, similarities, and time evolution. **International Journal of Educational Technology in Higher Education**, v. 12, n. 3, p. 98-112, 2015.

CHU, Minh Cong et al. Assessment of motorcycle ownership, use, and potential changes due to transportation policies in Ho Chi Minh City, Vietnam. **Journal of Transportation Engineering, Part A: Systems**, v. 145, n. 12, p. 05019007, 2019.

CROMPTON, Helen & BURKE, Diane. Artificial intelligence in higher education: the state of the field. **International Journal of Educational Technology in Higher Education.** 20. 10.1186/s41239-023-00392-8, 2023.

DA ROCHA ARANA, Andressa Maria Freire et al. Educational management and school effectiveness: a case study in municipal schools of Duque de Caxias. **Brazilian Journal of Development**, v. 4, n. 4, p. 1156-1167, 2018.

DA SILVA KOGLIN, Terena Souza; DE OLIVEIRA KOGLIN, João Carlos. A importância da extensão nas universidades brasileiras e a transição do reconhecimento ao descaso. **Revista Brasileira de Extensão Universitária**, v. 10, n. 2, p. 71-78, 2019.

Data Mining Fruitful and Fun. ORANGE, 2023. Disponível em: <https://orangedatamining.com/>. Acesso em: 16/06/2023.

DE LA CRUZ-CAMPOS, Juan-Carlos et al. Causes of academic dropout in higher education in Andalusia and proposals for its prevention at university: **A systematic review. In: Frontiers in Education. Frontiers,** 2023. p. 106.

DE MEDEIROS JÚNIOR, José Gilberto B.; FERRERO, Carlos Andrés. Agrupamento de Imagens Tumorais de MRI utilizando Extração de Descritores baseados em Séries Temporais. **In: Anais do XVII Escola Regional de Banco de Dados.** SBC, 2022. p. 109-118.

DE SOUSA, Angélica Silva; DE OLIVEIRA, Guilherme Saramago; ALVES, Laís Hilário. **A pesquisa bibliográfica: princípios e fundamentos. Cadernos da FUCAMP**, v. 20, n. 43, 2021.

DIMIC, Gabrijela et al. Improving the prediction accuracy in blended learning environment using synthetic minority oversampling technique. **Information Discovery and Delivery**, 2019.

DO NASCIMENTO, Rafaella Leandra Souza; DA CRUZ JUNIOR, Geraldo Gomes; DE ARAÚJO FAGUNDES, Roberta Andrade. Mineração de dados educacionais: um

estudo sobre indicadores da educação em bases de dados do INEP. **RENOTE**, v. 16, n. 1, 2018.

DOS SANTOS, Guilherme Mendes Tomaz, *et al*., Educação superior: reflexões a partir do advento da pandemia da COVID-19. **Boletim de conjuntura (BOCA)**, v. 4, n. 10, p. 108-114, 2020.

DOL, Sunita M.; JAWANDHIYA, Pradeep M. Use of Data mining Tools in Educational Data Mining. In: **2022 Fifth International Conference on Computational Intelligence and Communication Technologies (CCICT)**. IEEE, 2022. p. 380-387.

DUTTA, Shawni; BANDYOPADHYAY, Samir Kumar. Forecasting of Campus Placement for Students Using Ensemble Voting Classifier. **Asian J. Res. Comput. Sci**, v. 5, p. 1-12, 2020.

DUTT, Ashish; ISMAIL, Maizatul Akmar; HERAWAN, Tutut. A systematic review on educational data mining. **Ieee Access**, v. 5, p. 15991-16005, 2017.

DU, Xu et al. Educational data mining: a systematic review of research and emerging trends. **Information Discovery and Delivery**, v. 48, n. 4, p. 225-236, 2020.

EATHER, Narelle et al. Programmes targeting student retention/success and satisfaction/experience in higher education: **A systematic review. Journal of Higher Education Policy and Management**, v. 44, n. 3, p. 223-239, 2022.

FAISAL, M. et al. Comparative analysis of inter-centroid K-Means performance using euclidean distance, canberra distance and manhattan distance. In: **Journal of Physics: Conference Series. IOP Publishing**, 2020. p. 012112.

FUENTEALBA, Diego; LÓPEZ, Mario; PONCE, Héctor. Effects on time and quality of short text clustering during real-time presentations. **IEEE Latin America Transactions**, v. 19, n. 8, p. 1391-1399, 2021.

GOLDSMITH, Laurie J. Using Framework Analysis in Applied Qualitative Research. Qualitative Report, v. 26, n. 6, 2021.

GOVENDER, Paulene; SIVAKUMAR, Venkataraman. Application of k-means and hierarchical clustering techniques for analysis of air pollution: **A review (1980–2019). Atmospheric pollution research**, v. 11, n. 1, p. 40-56, 2020.

GUANDALINE, Valter Hugo. HCAIM: Um Método de Discretização Supervisionado para o Contexto de Classificação Hierárquica. 2016. **Tese de Doutorado. Universidade Federal de Ouro Preto**.

GUPTA, Satinder Bal; YADAV, Raj Kumar; GUPTA, Shivani. Analysis of Popular Techniques Used in Educational Data Mining. **International Journal of Next-Generation Computing**, p. 137-162, 2020.

GUSSO, Hélder Lima, *et al*., Ensino superior em tempos de pandemia: diretrizes à gestão universitária. **Educação & Sociedade**, v. 41, 2020.

HASAN, Raza, *et al*., Predicting student performance in higher educational institutions using video learning analytics and data mining techniques. **Applied Sciences**, v. 10, n. 11, p. 3894, 2020.

HERNÁNDEZ-BLANCO, Antonio, *et al*., A systematic review of deep learning approaches to educational data mining. **Complexity**, v. 2019, 2019.

HOLMES, Wayne; BIALIK, Maya; FADEL, Charles. Artificial Intelligence in Education. 2020.

HUNG, Hui-Chun, *et al.,* Applying educational data mining to explore students' learning patterns in the flipped learning approach for coding education. **Symmetry**, v. 12, n. 2, p. 213, 2020.

HUNG, Jui-Long et al. Improving predictive modeling for at-risk student identification: A multistage approach. **IEEE Transactions on Learning Technologies**, v. 12, n. 2, p. 148-157, 2019.

HUNTER, David; MCCALLUM, Jacqueline; HOWES, Dora. Defining exploratory-descriptive qualitative (EDQ) research and considering its application to healthcare. Journal of Nursing and Health Care, v. 4, n. 1, 2019.

INJADAT, MohammadNoor, *et al*., Systematic ensemble model selection approach for educational data mining. **Knowledge-Based Systems**, v. 200, p. 105992, 2020.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA (INEP). **Censo Escolar**, 2010. Brasília: MEC, 2011. JANUZZI, Paulo.

JAFARZADEGAN, Mohammad; SAFI-ESFAHANI, Faramarz; BEHESHTI, Zahra. Combining hierarchical clustering approaches using the PCA method. **Expert Systems with Applications,** v. 137, p. 1-10, 2019.

JALOTA, Chitra; AGRAWAL, Rashmi. Analysis of educational data mining using classification. In: **2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)**. IEEE, 2019. p. 243-247.

KAYE, Neil. Evaluating the role of bursaries in widening participation in higher education: **a review of the literature and evidence. Educational Review**, v. 73, n. 6, p. 775-797, 2021.

KAVIYA, K.; PREMALATHA, J. Prediction of compressive strength of high performance concrete using Artificial Neural Network (ANN) models. **International Research Journal of Engineering and Technology**, v. 6, p. 1378-1387, 2019.

KHAN, Anupam; GHOSH, Soumya K. Student performance analysis and prediction in classroom learning: A review of educational data mining studies. **Education and information technologies**, v. 26, n. 1, p. 205-240, 2021.

KUBRUSLY, Marcos; COELHO, Raquel Autran.; AUGUSTO, Kristopherson Lustosa; JUNIOR, Arnaldo Aires Peixoto; SANTOS, Daniela Costa de Oliveira; OLIVEIRA, Claudia Maria Costa de. Percepção docente sobre a Aprendizagem Baseada em Problemas no ensino remoto durante a pandemia COVID-19. **Research, Society and Development**, v. 10, n. 5, p. e53510515280-e53510515280, 2021.

LAZUARDI, Muhammad Lutfi; SUKOCO, Iwan. Design Thinking David Kelley & Tim Brown: Otak Dibalik Penciptaan Aplikasi Gojek. Organum: **Jurnal Saintifik Manajemen dan Akuntansi**, v. 2, n. 1, p. 1-11, 2019.

LORENZO-QUILES, Oswaldo; GALDÓN-LÓPEZ, Samuel; LENDÍNEZ-TURÓN, Ana. Dropout at university. Variables involved on it. In: **Frontiers in Education. Frontiers**, 2023. p. 124.

LUAN, Hui; TSAI, Chin-Chung. A review of using machine learning approaches for precision education. **Educational Technology & Society**, v. 24, n. 1, p. 250-266, 2021.

MACÊDO, Pedro HR; SANTOS, Wylliams B.; MACIEL, Alexandre MA. Análise de perfis de engajamento de estudantes de ensino a distância. **RENOTE**, v. 18, n. 2, p. 326-335, 2020.

MAHDI, Alyaa A. Educational Data Mining to Improve the Academic Performance in Higher Education. **Cihan University-Erbil Scientific Journal**, v. 4, n. 2, p. 13-18, 2020.

MASETHE, Mosima Anna, *et al.*, Framework of recommendation systems for educational data mining (edm) methods: Cbr-rs with knn implementation. In: **Transactions on Engineering Technologies**. Springer, Singapore, 2021. p. 87-98.

MAHAJAN, Ginika; SAINI, Bhavna. Educational Data Mining: A state-of-the-art survey on tools and techniques used in EDM. **International Journal of Computer Applications & Information Technology**, v. 12, n. 1, p. 310-316, 2020.

MAJERNÍK, Jaroslav et al. Development and implementation of an online platform for curriculum mapping in medical education. **Bio-Algorithms and Med-Systems**, v. 18, n. 1, p. 1-11, 2022.

Make sense of data, together. KNIME, 2023. Disponível em: <//www.knime.com/>. Acesso em: 16/06/2023.

MANJARRES, Andrés Villanueva; SANDOVAL, Luis Gabriel Moreno; SUÁREZ, Martha Salinas. Data mining techniques applied in educational environments: Literature review. **Digital Education Review**, n. 33, p. 235-266, 2018.

MEHRAD, Aida; ZANGENEH, Mohammad Hossein Tahriri. Comparison between qualitative and quantitative research approaches: Social sciences. **International Journal For Research In Educational Studies, Iran**, v. 5, n. 7, p. 1-7, 2019.

MISSAGLIA, Nelson. **Arquitetura para conexão de Ambiente Virtual De Aprendizagem com Tecnologias Inteligentes: Interpretação Automática das Interações Do Aprendiz.** Tese (Doutor em Informática e Gestão do Conhecimento) – Programa de Pós-Graduação em Informática e Gestão do Conhecimento da Universidade Nove de Julho - UNINOVE. 2019.

NEVES, Clarissa Eckert Baeta; HERINGER, Rosana; SAMPAIO, Helena. A institucionalização da pesquisa sobre ensino superior no Brasil. **Revista Brasileira de Sociologia**, v. 6, n. 12, p. 19-41, 2018.

NOETZOLD, Edson; PERTILE, Solange de L. Análise e predição de evasão dos alunos de um curso de Graduação em Sistemas de Informação por meio da mineração de dados educacionais. **RENOTE**, v. 19, n. 1, p. 351-360, 2021.

NOGUEIRA, Renan França Gomes; ALMEIDA, Luís Fernando. Proposta arquitetural de baixo custo para automação na coleta de dados de monitoramento de laboratórios em ambiente educacional baseado em internet das coisas (IoT). **Revista H-TEC Humanidades e Tecnologia**, v. 4, n. 1, p. 137-161, 2020.

OBERTO, Gonzalo Isaac Carrasco. **Cluster no jerárquicos versus cart y biplot**. 2020. Tese de Doutorado. Universidad de Salamanca.

PEDRO, Francesc et al. Artificial intelligence in education: **Challenges and opportunities for sustainable development.** 2019.

PEÑA-AYALA, Alejandro. Educational data mining: A survey and a data mining-based analysis of recent works. **Expert systems with applications**, v. 41, n. 4, p. 1432-1462, 2014.

PINTO, Neila Newdirley Câmara; MOTA, Sheila Cordeiro. Prototipação e validação multifásica de instrumento avaliativo para ensino de jovens e adultos| Prototyping and multiphasic validation of an evaluation tool for youth and adult education. **InfoDesign-Revista Brasileira de Design da Informação**, v. 17, n. 2, p. 49-65, 2020.

RAHMAN, Atta Ur et al. Unsupervised machine learning based documents clustering in Urdu. **EAI Endorsed Transactions on Scalable Information Systems**, v. 5, n. 19, p. e5-e5, 2018.

ROMERO, Cristobal; VENTURA, Sebastian. Data mining in education. **Wiley Interdisciplinary Reviews: Data mining and knowledge discovery**, v. 3, n. 1, p. 12-27, 2013.

RAMOS, Vanessa das Graças Santos, *et al*., Uma proposta de utilização de gestão de risco para o planejamento acadêmico de uma universidade pública. **Revista de Gestão e Projetos**, v. 10, n. 1, p. 81-91, 2019.

RapidMiner for Academics. RAPIDMINER, 2023. Disponível em: <https://rapidminer.com/platform/educational/>. Acesso em: 16/06/2023.

ROGERS, Frank. Educational fuzzy data-sets and data mining in a linear fuzzy real environment. **Journal of Honai Math**, v. 2, n. 2, p. 77-84, 2019.

RUIZ, Cristiane Regina. Criação de um modelo Canvas para planejamento acadêmico aliado a ferramentas de Design Thinking. **Revista on line de Política e Gestão Educacional**, p. 321-327, 2019.

SALAS-PILCO, Sdenka Zobeida; YANG, Yuqin. Artificial intelligence applications in Latin American higher education: a systematic review. **International Journal of Educational Technology in Higher Education**, v. 19, n. 1, p. 1-20, 2022.

SALAS-PILCO, Sdenka Zobeida; YANG, Yuqin; ZHANG, Zhe. Student engagement in online learning in Latin American higher education during the COVID-19 pandemic: **A systematic review. British Journal of Educational Technology**, v. 53, n. 3, p. 593-619, 2022.

SANDHYA, N.; RAMA PRASATH, A. Application of Artificial Intelligence Methods for Detection of Fronto Temporal Dementia. In: **International Conference on Intelligent Computing and Communication Technologies**. Springer, Singapore, 2019. p. 673-679.

SANTOS, Carolina da Costa; PEREIRA, Fátima; LOPES, Amélia. Experiências da gestão acadêmica da docência universitária. **Educação & Realidade**, v. 43, p. 989-1008, 2018.

SCHAFER, Dov; KAUFMAN, David. Augmenting Reality with Intelligent Interfaces. **Artificial intelligence: Emerging trends and applications**, v. 221, 2018.

SHARMA, Sushil Kumar; PALVIA, Shailendra C. Jain; KUMAR, Kuldeep. Changing the landscape of higher education: From standardized learning to customized learning. **Journal of Information Technology Case and Application Research**, v. 19, n. 2, p. 75-80, 2017.

SHELTON, Brett E.; HUNG, Jui-Long; LOWENTHAL, Patrick R. Predicting student success by modeling student interaction in asynchronous online courses. **Distance Education**, v. 38, n. 1, p. 59-69, 2017.

SHCHERBAN, Sergei et al. Multiclass Classification of UML Diagrams from Images Using Deep Learning. **International Journal of Software Engineering and Knowledge Engineering,** v. 31, n. 11n12, p. 1683-1698, 2021.

SIEMENS, George; BAKER, Ryan SJ d. Learning analytics and educational data mining: towards communication and collaboration. In: **Proceedings of the 2nd international conference on learning analytics and knowledge**. 2012. p. 252-254.

SILVA, Carla; FONSECA, José. Educational Data Mining: a literature review. **Europe and MENA Cooperation advances in information and communication technologies**, p. 87-94, 2017.

SIMÕES, Mara Leite. O surgimento das universidades no mundo e sua importância para o contexto da formação docente. **Universidade Federal da Paraíba. Revista Temas em Educação**, v. 22, n. 2, p. 136, 2013.

SINAGA, Kristina P.; YANG, Miin-Shen. Unsupervised K-means clustering algorithm. **IEEE access**, v. 8, p. 80716-80727, 2020.

Software IBM SPSS. SPSS, 2023. Disponível em: <https://www.ibm.com/br-pt/spss>. Acesso em: 16/06/2023.

SREENIVASA RAO, K.; SWAPNA, N.; PRAVEEN KUMAR, P. Educational data mining for student placement prediction using machine learning algorithms. **International Journal of Engineering and Technology (UAE)**, v. 7, n. 1.2, p. 43-46, 2018.

STOLL, Bruno Bastos. **Framework para Análise e Intervenção no Processo de Aprendizado.** Dissertação (Mestre em Informática) – o Programa de Pós-Graduação em Informática do Departamento de Informática da Universidade Federal do Espírito Santo. 2019.

STOLL, Bruno Bastos et al. Análise de dados acadêmicos baseado em previsão, recomendação e visualização. **RENOTE**, v. 17, n. 1, p. 286-295, 2019.

SUAVE, Ricardo; ALTOÉ, Stella Maris Lima; FERREIRA, Marcelo Marchine. Pesquisas experimentais aplicadas à educação contábil: **panorama atual e oportunidades no cenário brasileiro. Revista Contemporânea de Contabilidade**, v. 18, n. 47, p. 155-176, 2021.

TANG, Hengtao; XING, Wanli; PEI, Bo. Time really matters: Understanding the temporal dimension of online learning using educational data mining. **Journal of Educational Computing Research**, v. 57, n. 5, p. 1326-1347, 2019.

Unified engine for large-scale data analytics. APACHE SPARK, 2023. Disponível em: <https://spark.apache.org/>. Acesso em: 16/06/2023.

ULTSCH, Alfred; LÖTSCH, Jörn. Euclidean distance-optimized data transformation for cluster analysis in biomedical data (EDOtrans). **BMC bioinformatics,** v. 23, n. 1, p. 1-18, 2022.

VIVIAN-GRIFFITHS, Timothy, *et al*. Predictive modeling of schizophrenia from genomic data: Comparison of polygenic risk score with kernel support vector machines approach. **American Journal of Medical Genetics Part B: Neuropsychiatric Genetics**, v. 180, n. 1, p. 80-85, 2019.

YAĞCI, Mustafa. Educational data mining: prediction of students' academic performance using machine learning algorithms. **Smart Learning Environments**, v. 9, n. 1, p. 11, 2022.

ZAWACKI-RICHTER, Olaf; MARÍN, Victoria. I.; BOND, Melissa; GOUVERNEUR, Franziska. Systematic review of research on artificial intelligence applications in higher education–where are the educators?. **International Journal of Educational Technology in Higher Education**, v. 16, n. 1, p. 1-27, 2019.