ESCOLA POLITÉCNICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO
MESTRADO EM CIÊNCIA DA COMPUTAÇÃO

FERNANDO LUNARDELLI

ANÁLISE VISUAL DO PERCURSO ACADÊMICO DE ESTUDANTES AO LONGO DO ENSINO SUPERIOR

Porto Alegre
2022

PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL
ESCOLA POLITÉCNICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

ANÁLISE VISUAL DO PERCURSO ACADÊMICO DE ESTUDANTES AO LONGO DO ENSINO SUPERIOR

FERNANDO LUNARDELLI

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em Ciência da Computação na Pontifícia Universidade Católica do Rio Grande do Sul.

Orientadora: Profª. Isabel Harb Manssour

Porto Alegre
2022

Ficha Catalográfica

L961a Lunardelli, Fernando

Análise visual do percurso acadêmico de estudantes ao longo do ensino superior / Fernando Lunardelli. – 2022.
108.
Dissertação (Mestrado) – Programa de Pós-Graduação em Ciência da Computação, PUCRS.

Orientadora: Profa. Dra. Isabel Harb Manssour.

1. Análise Visual. 2. Ensino Superior. 3. Percurso Acadêmico. 4. Diagrama de Sankey. 5. Evasão. I. Harb Manssour, Isabel. II. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da PUCRS
com os dados fornecidos pelo(a) autor(a).
Bibliotecária responsável: Loiva Duarte Novak CRB-10/2079

FERNANDO LUNARDELLI

ANÁLISE VISUAL DO PERCURSO ACADÊMICO DE ESTUDANTES AO LONGO DO ENSINO SUPERIOR

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre em Ciência da Computação do Programa de Pós-Graduação em Ciência da Computação, Escola Politécnica da Pontifícia Universidade Católica do Rio Grande do Sul.

Aprovado(a) em 29 de Abril de 2022.

BANCA EXAMINADORA:

Profª. Drª. Milene Selbach Silveira (PPGCC/PUCRS)

Prof. Dr. Renato Perez Ribas (PGMicro/UFRGS)

Profª. Isabel Harb Manssour (PPGCC/PUCRS - Orientadora)

DEDICATÓRIA

Dedico este trabalho à minha esposa Cláudia e ao meu filho João Lucca, por todo o imenso apoio ao longo dos muitos alvoreceres de trabalho e isolamento social, e em especial, à minha mãe Nair, por sempre apoiar e incentivar meus estudos.

“The obvious is that which is never seen until someone expresses it simply.”
(Khalil Gibran)

AGRADECIMENTOS

Quero iniciar com um super, hiper, mega, ultra, blaster, agradecimento à minha orientadora Isabel H. Manssour, por acreditar neste trabalho e na pessoa que o conduziu. Obrigado pelas inúmeras horas dedicadas às correções e sugestões para que este trabalho se tornasse menos imperfeito.

Agradeço também por todas as correções e importantes comentários feitos pela professora Milene Selbach, avaliadora deste trabalho ao longo de todos os seus *milestones*.

À UOL Edutech, por proporcionarem a bolsa de estudos utilizada neste trabalho.

À PUCRS, ao programa de Pós-Graduação em Ciência da Computação e a todos os seus professores e colegas, por todo o apoio e recursos oferecidos (saudades dos muitos grupos de estudos que fizemos).

Agradeço aos profissionais entrevistados, pelo seu tempo e valiosos “insights” proporcionados. Espero que possamos conversar outras vezes sobre o tema.

Ao meu irmão Fabrício, sua esposa Mei e meus sobrinhos, Vicente e Gabriel, pelo apoio constante e reflexões.

Aos amigos Marcelo e Flávia Azambuja, por me incentivarem e apoiarem na realização deste sonho de mestrado acadêmico pela PUCRS. Se não fosse por eles, possivelmente, *Piaget* estaria sendo citado outras vezes neste texto.

Aos amigos e colegas do grupo de pesquisa DaVInt, e em especial, à amiga Daiane, pelas orientações e apoio ao longo da criação do modelo preditivo.

O amigo Leonardo S., por se empenhar na disponibilização dos dados utilizados neste estudo.

Ao amigo Felipe “Fido”, pelas considerações acadêmicas e apoio ao longo da pesquisa.

Finalmente, agora que escrevo as últimas palavras deste trabalho (ao menos antes das correções e sugestões que virão), paro e reflito no que me aconteceu em âmbito pessoal, ao longo deste período de mestrado (duas mudanças de emprego, mudança de moradia, um filho e uma pandemia), toda a dedicação requerida por este curso, dos muitos fins de semana e horas de estudo, dos grupos estudo e trocas com colegas e professores, enfim, de todo o esforço necessário para este momento. Chego a conclusão de

que tudo isso fez parte de uma jornada incrível de (auto-)conhecimento e agradeço por todos esses momentos.

“This is the way”[1].

[1] Citação Mandaloriana

ANÁLISE VISUAL DO PERCURSO ACADÊMICO DE ESTUDANTES AO LONGO DO ENSINO SUPERIOR

RESUMO

De acordo com o Censo da Educação Superior no Brasil, a evasão nos cursos de graduação é um problema que está piorando a cada ano nas instituições de ensino superior. Porém, uma análise constante e unificada do percurso do aluno pode ajudar a melhorar este cenário, auxiliando a entender ou prever quando não haverá a conclusão destes cursos. Entretanto, para isso, são necessárias ferramentas analíticas que facilitem estes acompanhamentos e viabilizem a tomada de decisões. Neste contexto, o presente trabalho propõe a criação de um modelo de visualização de dados que possibilite a análise do percurso acadêmico de um ou mais alunos durante o ensino superior. Através da exploração de dados e análise estatística, este modelo visa permitir a identificação de indivíduos, ou grupos de indivíduos, com tendência a não completarem seus cursos com sucesso, além de permitir uma "visão do todo" em relação ao seu percurso acadêmico e principais indicadores. Desta forma, busca auxiliar os tomadores de decisão das instituições de ensino (administradores, educadores, responsáveis técnicos, etc.), na condução de orientações, aplicação de políticas e outras ações, que minimizem as condições que levam estes alunos à evasão. O modelo proposto, centrado em uma visualização que utiliza diagrama de *Sankey*, conectado a um modelo de predição de evasão, e sua implementação, foram baseados nos requisitos identificados a partir de uma revisão sistemática da literatura, da implementação de um protótipo e de entrevistas com quatro especialistas de domínio. A implementação do modelo também foi validada através de entrevistas com quatro especialistas de domínio, que a consideraram adequada à contribuir para a melhora do acompanhamento de progresso estudantil.

Palavras-Chave: análise visual, ensino superior, percurso acadêmico, diagrama de sankey, evasão, modelo de predição.

VISUAL ANALYTICS OF THE ACADEMIC PATH OF STUDENTS IN HIGHER EDUCATION

ABSTRACT

According to the Census of Higher Education in Brazil, dropout in undergraduate courses is a problem getting worse every year in higher education institutions. However, constant and unified analysis of the student's path can help improve this scenario, enabling understanding or predicting when these courses will not be completed. Nonetheless, analytical tools are needed to facilitate these follow-ups and make decision-making feasible. In this context, the present work proposes creating a data visualization model that allows the analysis of the academic path of one or more students during higher education. Through the exploration of data and statistical analysis, this model aims to identify individuals, or groups of individuals, with a tendency to not complete their courses successfully, in addition to allowing a "view of the whole" concerning their academic career and key indicators. In this way, it seeks to help decision-makers of educational institutions (administrators, educators, technical managers, etc.), in conducting guidelines, applying policies, and other actions, which minimize the conditions that lead these students to drop out. The proposed model, centered on a visualization that uses a *Sankey* diagram connected to an evasion prediction model, and its implementation, were based on the requirements identified from a systematic literature review, the implementation of a prototype, and interviews with four domain experts. The implementation of the model was also validated through interviews with four domain experts, who considered it adequate to contribute to the improvement of student progress monitoring.

Keywords: visual analytics, higher education, academic path, sankey diagram, dropout, prediction model.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

API – *Application Programming Interface* - Interface de Programação de Aplicativos

AVA – Ambientes Virtuais de Aprendizagem

CAAE – Certificado de Apresentação de Apreciação Ética

CRISP-EDM – *CRoss-Industry Standard Process for Educational Data Mining* - Processo Padrão Inter-Indústrias para Mineração de Dados Educacionais

CSV – *Comma-Separated Values* - Valores Separados por Vírgula

DAG – *Directed Acyclic Graph* - Grafos Acíclicos Dirigidos

EDA – Exploratory Data Analysis - Análise Exploratória de Dados

EDM – *Educational Data Mining* - Mineração de Dados Educacionais

GPA – *Grade Point Average* - Média de Pontos de Classificação

LA – *Learning Analytics* - Análise de Aprendizado

QP – Questão de Pesquisa

RSL – Revisão Sistemática da Literatura

SVG – *Scalable Vector Graphic* - Gráfico de Vetor Escalável

TCLE – Termo de Consentimento Livre e Esclarecido

VIF – *Variance Inflation Factor* - Fator de Inflação da Variância

ROC – *Receiver Operating Characteristic* - Curva Característica de Operação do Receptor

ROC-AUC – *Receiver Operating Characteristic - Area Under Curve* - Característica de Operação do Receptor - Área sob a curva

SUMÁRIO

1. INTRODUÇÃO

De acordo com os Censos da Educação Superior no Brasil [34], a evasão nos cursos de graduação é um problema notório nas instituições de ensino superior. O total de alunos que concluem seus cursos não tem acompanhado o número de ingressos ao longo dos anos. No censo de 2020, 3.765.475 estudantes entraram no ensino superior, mas somente 1.278.622 (33.9%) concluíram seus cursos. Este índice é inferior aos indicadores de 2019 (34.4%) ou mesmo 2017 (37%), levando ao entendimento que, esta diferença tende a se agravar caso ações não sejam tomadas. Além disso, este indicador configura uma forte necessidade de uma análise aprofundada sobre o percurso estudantil no ensino superior brasileiro [12]. Portanto, identificar alunos com um grande atraso na formação [28], com risco de falharem ou desistirem do curso é muito relevante para os tomadores de decisão das instituições de ensino (administradores, educadores, responsáveis técnicos, etc.), porém, nem sempre estes possuem as ferramentas, dados e preparo adequados para estas análises [6]. Informações como cursos com maior procura ou a situação de cada aluno nas disciplinas são acompanhadas de forma individual e reativa, por demanda, não sendo possível, segundo Raji et al. [53], “ter uma visão do ‘todo’” de forma rápida e fácil. A dificuldade em gerenciar, manipular e processar este grande volume de dados temporais, obtidos ao longo dos semestres e relacionados a este contexto, torna uma simples tarefa de análise excessivamente laboriosa sem a utilização de ferramentas especializadas [31, 24]. Esta “falta analítica” pode limitar ações importantes, como, por exemplo, uma análise mais profunda das razões pelas quais um aluno está atrasado em relação ao andamento esperado, entre outros *insights* que podem repercutir em uma ação de orientação, levando-o a uma conclusão bem sucedida do curso [53].

O processo de descoberta de padrões úteis, a partir de grandes quantidades de dados de forma automática ou semiautomática [75], é conhecido como Mineração de Dados na literatura. Com enfoque educacional, uma subárea emergente, mas de grande relevância [58], *Educational Data Mining* (EDM), pode ser definida como um campo de investigação científica centrada no desenvolvimento de métodos de descoberta a partir dos dados originados de ambientes educacionais, com o objetivo de entender melhor os alunos e seus ambientes de aprendizagem [40].

Seguindo uma abordagem de esforços semelhantes mas com pequenas diferenças, a área de pesquisa chamada *Learning Analytics* (LA) [67], se concentra no processo de aprendizagem (que inclui a análise da relação entre aluno, conteúdo, instituição e educadores. Esta abordagem surge da necessidade de uma visão mais integrada e apurada dos dados de ensino, auxiliando tomadores de decisão a correlacionarem dados dos alunos à forma como interagem com os Ambientes Virtuais de Aprendizagem (AVA).

Outra área de pesquisa chamada de *Academic Analytics*, ou Análise Acadêmica, é descrita por Campbell et al. [13] como um mecanismo para a tomada decisões ou para orientar ações. Diferentemente de LA, *Academic Analytics* possui um enfoque na análise de dados em nível institucional, por meio de técnicas estatísticas e de modelagem preditiva. Suas ações, tanto podem incluir, análises em dados que antecedem uma matrícula, quanto dados históricos curriculares, visando auxiliar no acompanhamento de alunos que estão com dificuldades acadêmicas [13].

De forma a contribuir com esta Análise Acadêmica, a utilização do *Visual Analytics*, ou Análise Visual, descrita como a ciência do raciocínio analítico facilitado por interfaces visuais interativas [18], possibilita um olhar temporal e relacional acerca das movimentações dos estudantes sob diferentes perspectivas, estendendo a capacidade dos agentes nas instituições de ensino e permitindo uma visão ampliada de todo esse processo acadêmico [53]. Ao fornecer a capacidade de análise exploratória visual, rápida e intuitiva dos fluxos de alunos em várias coortes[1], é possível incentivar a interação e o pensamento crítico sobre os dados, a formulação de perguntas e os direcionamentos necessários para as ações práticas ou "*actionable insights*" [36].

A respeito de seu trabalho na área, Heilerman et al. (2015, p. 32) [31] comentam que, os recursos de análise visual provaram ser vitais para o contexto de acompanhamento acadêmico por dois motivos:

> Em primeiro lugar, essas ferramentas permitem a rápida exploração de grandes conjuntos de dados, fornecendo a capacidade de identificar tendências e anomalias mais facilmente nos dados relacionados ao progresso do aluno. Em segundo lugar, a capacidade de fornecer provas visuais do progresso do aluno, ou a falta do mesmo, rapidamente fundamentam discussões sobre o sucesso do aluno em fatos, em vez de especulações ou suposições.

Considerando este contexto, o objetivo deste trabalho é facilitar o acompanhamento do percurso acadêmico de estudantes ao longo do ensino superior, auxiliando os responsáveis na exploração dos dados institucionais através de um modelo de análise visual. O modelo proposto, centrado em uma visualização que utiliza diagrama de *Sankey* conectada a um modelo de predição de evasão, e sua implementação, foram baseados nos requisitos identificados a partir de uma revisão da literatura, da implementação de um protótipo e de entrevistas com quatro especialistas de domínio.

As principais contribuições deste trabalhos são:

- Uma revisão sistemática da literatura (RSL) sobre análise do percurso acadêmico ao longo do ensino superior, incluindo a descrição dos trabalhos selecionados e uma discussão sobre as técnicas de visualização utilizadas e suas contribuições para o cenário educacional.

[1] Coorte é uma amostra ou grupo, selecionada de uma parte ou mesmo da totalidade de uma população, e definida para avaliação longitudinal das relações exposição-resultado [70].

- Um conjunto de requisitos elicitados[2] a partir da RSL e de entrevistas com quatro profissionais especialistas de domínio, ou seja, que fazem o acompanhamento dos estudantes através da análise de dados.
- Um modelo para análise visual do percurso acadêmico de estudantes no ensino superior, contendo arranjos visuais para o detalhamento de informações de alunos e disciplinas, uma visão de progresso e um modelo preditivo.
- Uma implementação do modelo proposto utilizando uma visualização interativa baseada em diagrama de *Sankey* conectada a um modelo preditivo de evasão por regressão logística. Esta implementação do modelo foi analisada, também, através de entrevistas com quatro especialistas de domínio. Após interagirem com a implementação, os especialistas foram questionados sobre suas visualizações e funcionalidades, e foram unânimes em considerar que esta proposta pode contribuir para a melhora do acompanhamento estudantil.

O restante deste trabalho esta organizado da seguinte forma: O Capítulo 2 apresenta a metodologia utilizada para o desenvolvimento deste trabalho. O Capítulo 3 contém a descrição do processo de seleção e uma análise dos trabalhos relacionados, a partir dos resultados da revisão sistemática da literatura. Uma descrição e justificativa para a implementação de um protótipo para análise visual do percurso acadêmico de estudantes é apresentada no Capítulo 4. As entrevistas realizadas com especialistas de domínio para análise de requisitos são descritas no Capítulo 5. O modelo proposto com base na análise dos requisitos identificados e sua implementação são descritos em detalhes no Capítulo 6. Uma discussão sobre os resultados obtidos com o modelo, que foi apresentado à especialistas de domínio, é feita no Capítulo 7. Por fim, o último capítulo contém as conclusões finais do trabalho.

[2] Processo por meio do qual clientes e usuários são questionados sobre suas necessidades quanto ao software [71].

2. METODOLOGIA DE PESQUISA

Para alcançar o objetivo proposto e guiar o desenvolvimento deste trabalho, foi definida a seguinte questão de pesquisa (QP):

Como um modelo de análise visual poderia auxiliar no acompanhamento do percurso acadêmico dos estudantes?

Para responder a QP, o estudo foi organizado em duas etapas, apresentadas na Figura 2.1: *exploratória* e *execução*. A primeira abrange as atividades relacionadas à pesquisa exploratória, como definições iniciais (questões e objetivos), revisão sistemática da literatura, entrevistas semiestruturadas em profundidade com especialistas de domínio (com protocolo previamente enviado ao Comitê de Ética, na etapa de execução) para avaliação do protótipo, assim como a avaliação do modelo e implementação, posteriormente. A segunda diz respeito às atividades relacionadas ao desenvolvimento do protótipo de visualização, levantamento de requisitos, projeto e implementação do modelo (incluindo modelo preditivo) para análise do percurso acadêmico, além das contribuições e conclusão da pesquisa. Cada uma destas atividades é apresentada a seguir e descrita com mais profundidade nos próximos capítulos.

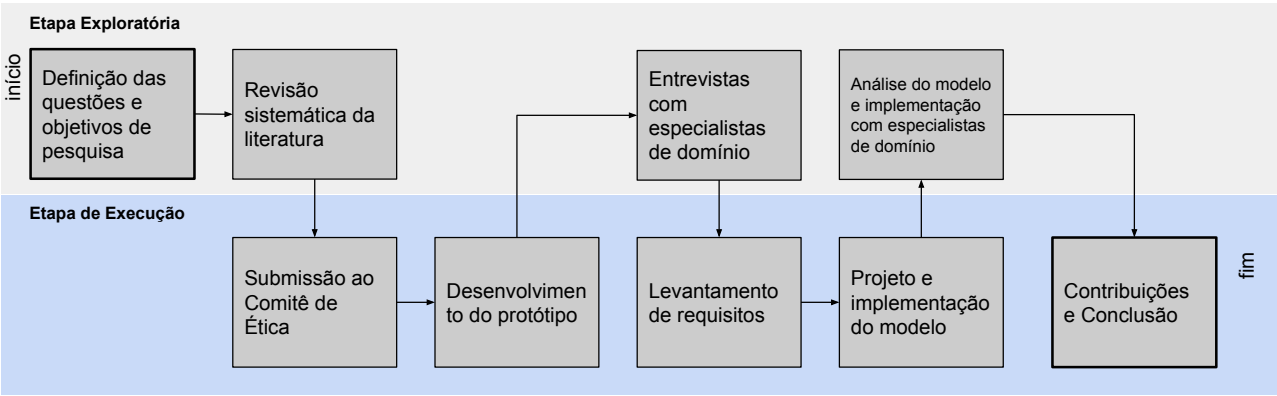


Figura 2.1: Atividades principais da pesquisa.

Inicialmente foram analisados os trabalhos relacionados descritos no Capítulo 3, a partir da condução de uma Revisão Sistemática da Literatura (RSL). Segundo Booth [9], uma RSL segue um passo a passo que minimiza erros sistemáticos e aleatórios na busca pelos trabalhos [20] e que pode ser seguido por outros que desejarem pesquisar sobre o mesmo tópico. Este tipo de revisão, auxilia no entendimento do assunto de interesse e a identificar o que já foi pesquisado a seu respeito e o que ainda pode ser explorado.

Portanto, a partir da RSL foi possível identificar um conjunto de trabalhos relacionados ao tema que foram descritos e avaliados criticamente. Assim, foi possível entender o atual contexto de acompanhamento do percurso acadêmico, ferramentas empregadas, técnicas de visualização, formas de interação e técnicas estatísticas utilizadas. O resultado desta análise auxiliou na identificação de oportunidades de pesquisa e levou à

definição e implementação de um protótipo de visualização de percurso acadêmico apresentado no Capítulo 4. Também neste etapa de protótipo foram buscadas bases de dados que pudessem auxiliar no desenvolvimento.

Para complementar a RSL e fazer um melhor levantamento de requisitos para este estudo, identificando também as necessidades e ferramentas utilizadas na prática, foram feitas entrevistas semiestruturadas em profundidade com profissionais especialistas de domínio. Para condução destas entrevistas, foi seguido o protocolo estabelecido pelo Comitê de Ética e Pesquisa[1], ou seja, as entrevistas foram realizadas somente após a sua aprovação por este comitê. A documentação com o protocolo de pesquisa pode ser encontrada no site da Plataforma Brasil[2], (número CAAE 29383420.0.0000.5336).

Além de questões sobre a forma como os especialistas fazem o acompanhamento dos estudantes, o protótipo implementado foi apresentado e analisado. O resultado destas entrevistas e o conjunto de requisitos identificados estão descritos no Capítulo 5.

Por meio do processo de elicitação de requisitos um modelo visual foi proposto e implementado, juntamente com um modelo de predição de evasão. Todas as funcionalidades e visualizações disponíveis estão descritas no Capítulo 6.

A implementação do modelo foi apresentada à especialistas de domínio em um novo processo de entrevistas, com o objetivo de analisar o modelo proposto e auxiliar a responder a QP. Estas entrevistas junto de uma discussão sobre as contribuições e limitações do trabalho são apresentadas no Capítulo 7. Por fim, as conclusões finais são apresentadas em detalhes no último Capítulo.

[1] https://www.pucrs.br/pesquisa/comites/cep
[2] https://plataformabrasil.saude.gov.br/login.jsf

3. TRABALHOS RELACIONADOS

Nesta seção são apresentados os trabalhos relacionados encontrados a partir de uma RSL que seguiu a metodologia proposta por Kitchenham [39]. Seguindo esta metodologia, foram definidas questões norteadoras para auxiliar na qualificação do processo de exploração da literatura, além de critérios de exclusão e inclusão dos trabalhos encontrados. De forma a complementar à RSL, foi realizada uma etapa de *forward snowballing* por meio da ferramenta de buscas *Google Scholar*[1]. Este método consiste em procurar a literatura que referencia os artigos selecionado para identificar novos trabalhos que potencialmente sejam relevantes para a pesquisa. O resultado desta etapa apresenta boa precisão e reduz consideravelmente o esforço na complementação da RSL [22]. Uma visão geral de todo processo é apresentada na Figura 3.1 e cada etapa é descrita a seguir.

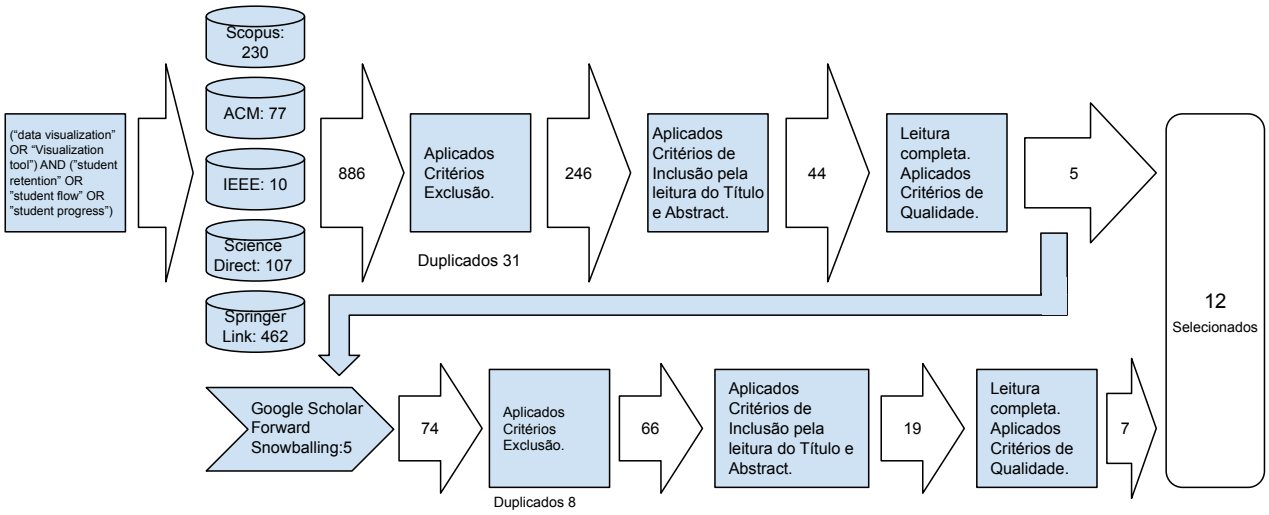


Figura 3.1: Processo de Revisão Sistemática da Literatura.

O processo de seleção de trabalhos foi precedido pela escolha das questões norteadoras que orientaram a revisão sistemática de literatura. Estas questões são apresentadas a seguir:

QP1: Como são visualizados os dados de fluxo e progresso dos alunos?

QP2: Por meio das visualizações apresentadas, quais resultados comuns foram encontradas na literatura, objetivando promover uma melhoria no cenário educacional?

[1]https://scholar.google.com

Após esta etapa, foi definida a expressão de busca apresentada na Tabela 3.1 e as bases digitais a serem consultadas: *Springer Link*[2], *ACM*[3] e *IEEE Digital Library*[4], *Elsevier ScienceDirect*[5] e *Elsevier Scopus*[6].

Tabela 3.1: Expressão de busca utilizada na RSL.

Expressão
(“data visualization” OR “visualization tool” OR “visual analytics”) AND (“student retention” OR “student flow” OR “student progress”)

Na sequência, foram definidos os seguintes critérios de exclusão:

- Artigos que não sejam de conferência, revisão[7] ou periódicos;
- Artigos que não sejam no idioma inglês;
- Artigos não disponíveis para leitura dentro dos acessos disponibilizados pela instituição;
- Artigos que não sejam da área de Ciência da Computação;
- Artigos que não tenham sido publicados entre Janeiro de 2010 até Janeiro de 2022;
- Artigos duplicados.

Os critérios de inclusão utilizados foram:

- Artigos que apresentem uma representação visual de dados;
- Artigos que tratam do contexto educacional.

Portanto, na etapa de qualificação, os estudos selecionados deveriam atender a estes critérios de inclusão e exclusão, além de responderem às questões QP1 e QP2.

Após a aplicação da expressão de busca nas bases selecionadas foram identificados 886 artigos, que após serem filtrados segundo os critérios de exclusão, resultaram em 246 artigos. Após a aplicação dos critérios de inclusão, por meio da leitura de título e *abstract*, foram selecionados 44 artigos aderentes ao tema de pesquisa. Depois da leitura completa destes artigos, aplicando os critérios de qualidade, ou seja, ajudam a responder QP1 e QP2, e devido a especificidade deste estudo, foram identificados apenas

[2] https://link.springer.com/
[3] https://dl.acm.org/
[4] https://ieeexplore.ieee.org/Xplore/home.jsp
[5] https://www.sciencedirect.com/
[6] https://www.scopus.com/search
[7] *Review Papers*

5 artigos completamente alinhados às questões (QP1 e QP2) e selecionados para a análise dos resultados. A partir desta primeira iteração os artigos aderentes foram utilizados no processo de *forward snowballing* no qual foram identificados 74 artigos que faziam referência aos 5 primeiros. Aplicados os critérios de exclusão e inclusão com a leitura do título e *abstract*, foram selecionados 19 artigos para leitura completa, sendo que o resultado final desta etapa foi de 7 artigos aderentes. Totalizando 12 artigos, que são apresentados na Seção 3.1. Na sequência, uma análise destes trabalhos, juntamente com as questões norteadoras, é apresentada na Seção 3.2.

3.1 Descrição dos Trabalhos

Nesta seção são apresentados os 12 trabalhos selecionados com uma breve descrição das suas propostas.

O trabalho de Heileman et al. [31] descreve um sistema desenvolvido para explorar o percurso acadêmico de alunos pela Universidade do Novo México (EUA) a partir do ano de 2007. Foram usados diagramas de *Sankey* para auxiliar na análise e "corrigir vários equívocos" relacionados ao sucesso dos alunos no campus.

De forma análoga à visualização de percurso acadêmico, também segundo Heileman et al. [31], diagramas de *Sankey* têm sido usados para visualização de percurso e fluxos, desde as suas primeiras aparições em meados de 1800. Tradicionalmente, são usados para a visualização de fluxos de energia, ilustrando informações quantitativas, seus relacionamentos, sua transformação e ajudando na indicação de ineficiências e potencial pontos de economia nas conexões de recursos [61].

O sistema proposto no trabalho permite, por meio dos diagramas, escolher uma divisão (faculdade) dentro da instituição e explorar como os alunos entram e saem dela e suas subdivisões (departamentos e cursos) ao longo do tempo, considerando aqueles que pararam ou se formaram ao longo do caminho. Detalhes de cada elemento que compõe o diagrama de *Sankey* podem ser vistos pela interação com mouse. Segundo os autores, a utilização destas visualizações interativas forneceram uma visão de alto nível, muito útil no direcionamento de investigações adicionais e que ajudou a corroborar com as discussões do campus sobre o sucesso dos alunos. A Figura 3.2 apresenta um fluxo de alunos de uma coorte de 2007 da Faculdade de Artes e Ciências. O diagrama é dividido em diferentes períodos (ou *Terms*) nos quais o fluxo de alunos é apresentado em destaque, como por exemplo: *OUTSIDE*, que representa alunos de outras faculdades ou outras subdivisões dentro das Artes e Ciências; *STOP*, que representa alunos que abandonaram seus cursos; e *GRAD*, representando alunos graduados. Segundo os autores, após seis anos, os dados mostram que apenas 6% dessa subpopulação continuou a se formar fora da di-

visão de ciências naturais, enquanto 10% ainda estavam matriculados (estendendo sua permanência na instituição), mas não conseguiram se formar.

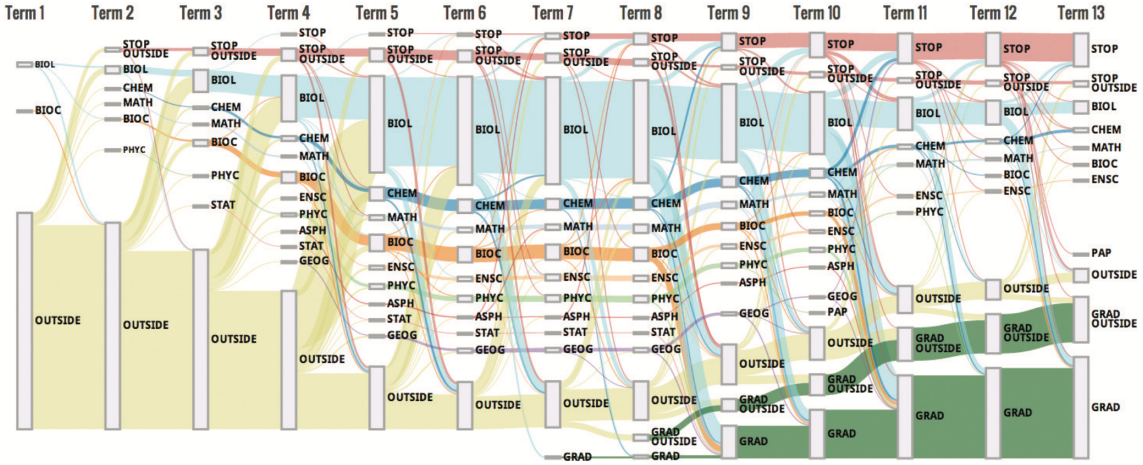


Figura 3.2: Exemplo do fluxo de alunos da coorte de 2007 que participaram dos programas de graduação da divisão de ciências naturais da faculdade de Artes e Ciências. Fonte: Heileman et al. [31]

Uma ferramenta de visualização de dados chamada *Ribbon Tool* é apresentada por Greer et al. [27] em um estudo de caso na UC Davis (EUA). A *Ribbon Tool* fornece uma visualização interativa dos fluxos de alunos através de programas acadêmicos e progresso ao longo do tempo, culminando na conclusão bem-sucedida (graduação) ou evasão. Junto com a visualização, os tomadores de decisão têm acesso a um conjunto de filtros que podem ser utilizados para aprimorar a exploração de dados, possibilitando comparações diversas, como, por exemplo, sobre o acompanhamento da efetividade de mudanças curriculares. Na Figura 3.3 é representado um diagrama de *Sankey* com três barras verticais indicando os agrupamentos de estudantes e os períodos de Setembro de 2011, de 2014 e de 2015. Na visualização, as cores das linhas representam agrupamentos de estudantes que se conectam aos possíveis estados do currículo: Matriculado (*Enrolled*), Parado (*Stopout*), Formado (*Awarded*) e Evadido (*Left*). Nesta mesma figura, por exemplo, as linhas vermelhas mostram o número de alunos que começaram em Engenharia (Setembro de 2011) e sua movimentação à medida que avançam no tempo.

Raji et al. [53] apresentam um sistema de descoberta de conhecimento visual chamado *eCamp* que reúne uma variedade de fontes de dados estudantis que antes estavam desconectadas em sistemas isolados. Os dados utilizados incluem notas de alunos, cursos principais e registros de graduação. O sistema é dividido em duas visualizações que estão ilustradas na Figura 3.4. A primeira, para análise do progresso do aluno, usa uma árvore radial (*radial tree*) com bordas adaptadas e semelhantes a um diagrama de *Sankey* para fornecer uma melhor percepção do fluxo de alunos ao longo do tempo. A segunda é uma visualização de sucesso do aluno, representada por um grafo (*node-link diagram*) composto por todas as áreas principais de um curso (*majors*). Este grafo utiliza

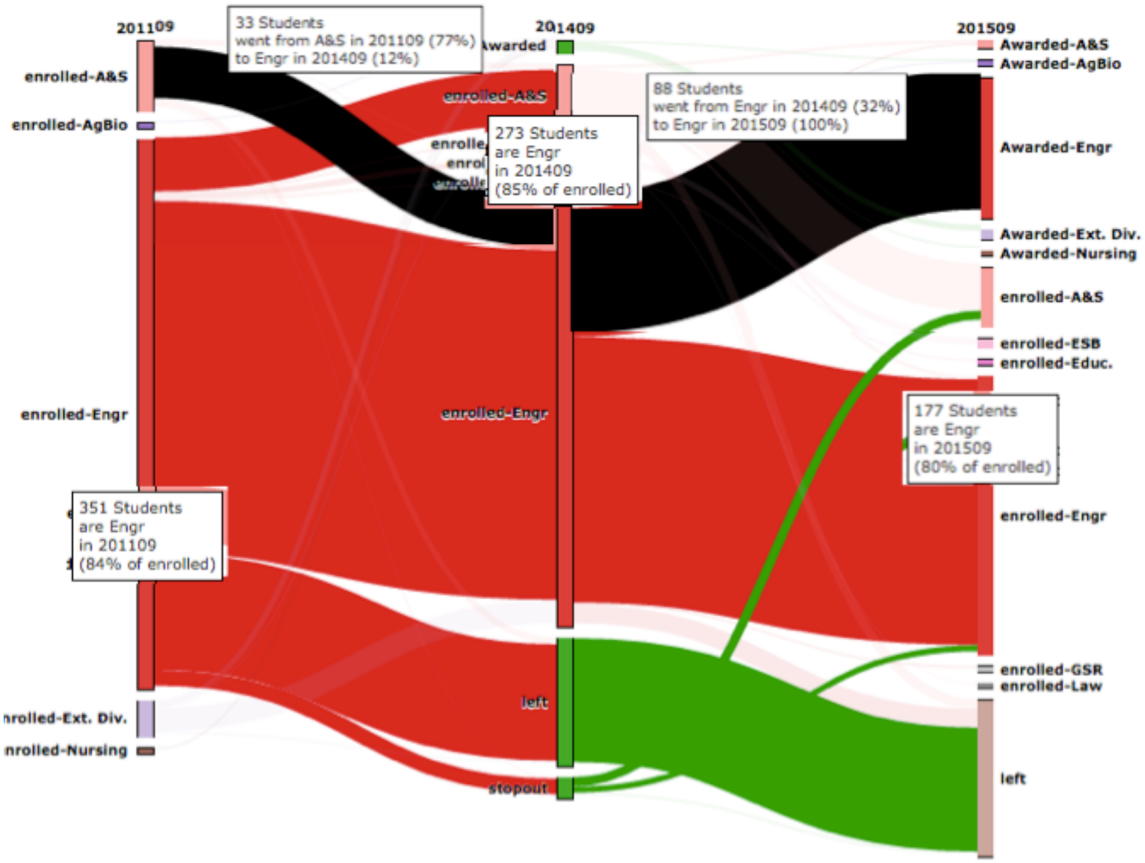


Figura 3.3: Exemplo de visualização da *Ribbon Tool* aplicada aos dados de estudantes de Engenharia no período de 2011 a 2015. Fonte: Greer et al. [27]

gráficos de pizza em que o tamanho representa o número de reprovações, enquanto as cores azul e vermelho, representam, respectivamente, alunos do sexo masculino e feminino.

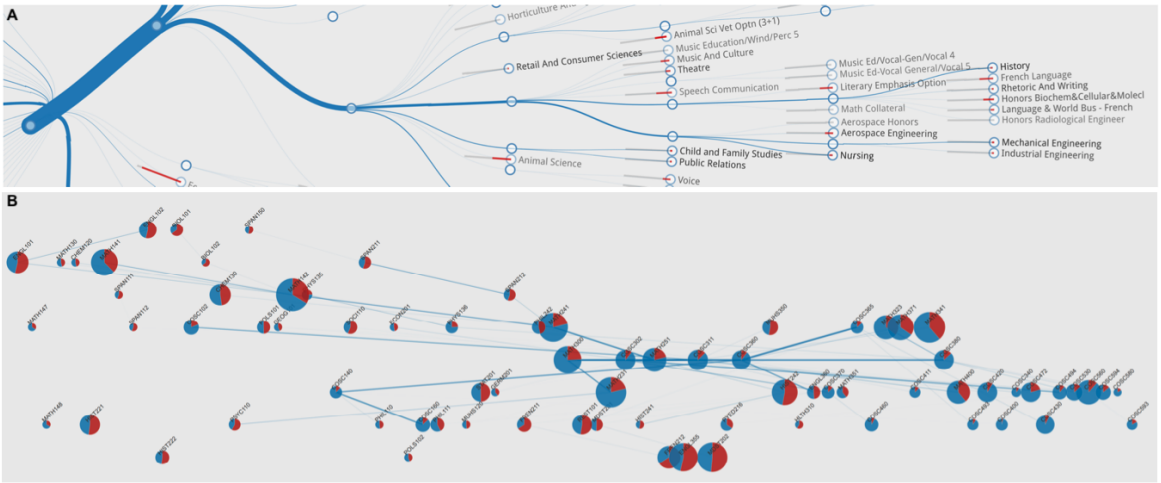


Figura 3.4: (A) Detalhamento do ramo do gráfico radial mostrando o caminho para a Engenharia Mecânica e a Engenharia Industrial. (B) Grafo conectando as áreas principais, indicando a proporção de alunos que falharam e o gênero. Fonte: Raji et al. [53]

Os autores Basavaraj et al. [8] procuram entender possíveis impactos no sucesso acadêmico de alunos da Universidade Central da Flórida (EUA), a partir da comparação de dois cursos de graduação diferentes - Ciência da Computação (CS) e Tecnologia da Informação (IT) sob um mesmo departamento, utilizando diagrama de *Sankey* para demonstrar o percurso acadêmico. O estudo analisou, entre 2008 e 2013, a relação destes dois cursos sob uma perspectiva estatística (quantitativa) e social (qualitativa) e concluiu que, sob uma ótica de indicadores de progresso e sucesso, existem poucas diferenças entre os cursos relacionadas a índices de evasão e graduação. Porém, uma análise qualitativa sugeriu uma relação hierárquica negativa entre os dois cursos, em que IT acaba sendo percebido como mais “fácil” quando comparado a CS, que é mais volátil e rigoroso em termos de percurso pelas áreas principais (*majors*), na percepção dos alunos. A Figura 3.5 apresenta um detalhamento da comparação dos dois cursos (CS e IT), na qual as colunas representam os períodos letivos e os nós que às compõem representam os demais cursos da instituição, assim como os possíveis destinos do estudante (matrícula, graduação, evasão). Na Figura Figura 3.5A, pode-se perceber o fluxo de alunos de IT, a partir do primeiro período (*Fall 2008*), movimentando-se para outros cursos (*Arts and humanities, Business Administration, Sciences e CS*), evadindo ou se graduando conforme avançam no tempo.

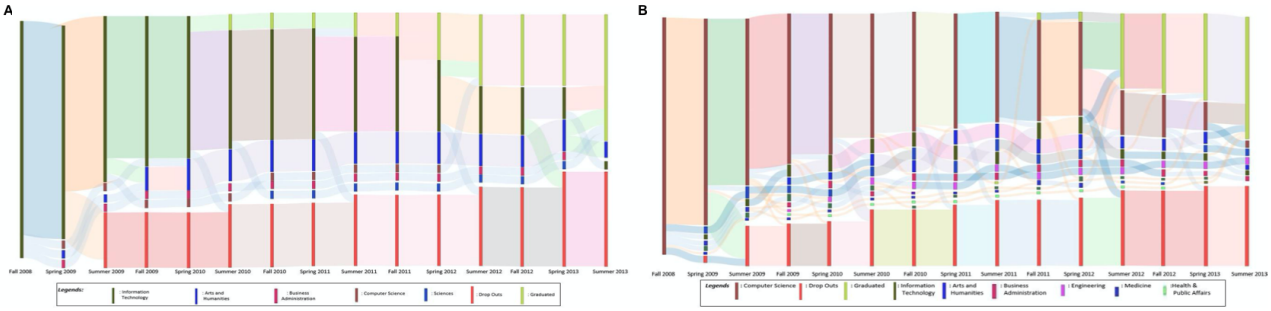


Figura 3.5: Detalhamento da comparação entre os cursos de Tecnologia da Informação (IT) (A) e Ciência da Computação (CS) (B). Fonte: Basavaraj et al. [8]

Um painel de análise de aprendizagem (*Learning Analytics Dashboard*) interativo baseado na web foi apresentado por Vaclavek et al. [72]. Este painel foi implantado com sucesso na Faculdade de Engenharia Mecânica (FME) da Universidade Técnica Tcheca em Praga. O sistema é composto por múltiplas visões analíticas (uma visão de progresso e gráfico de probabilidades de conclusão de créditos) e uma visão de resumo. A visão de progresso, ilustrada na Figura 3.6, contém um gráfico de linhas representando alunos divididos em cinco grupos de desempenho. Cada grupo é representado por uma linha que mostra a média de créditos alcançados pelos alunos para cada semana do ano letivo. Linhas pontilhadas dividem o gráfico em vários segmentos para melhor orientação ao

longo do ano acadêmico. No trabalho a estimativa de classificação em um dos grupos de desempenho é calculada usando o teorema de Bayes.

Figura 3.6: Visão geral de análise com filtro, menu de navegação e gráfico de progresso de estudos. Fonte: Vaclavek et al. [72]

Outra ferramenta de visualização é apresentada por Horvath et al. [32], para a análise de padrões de fluxo de alunos por diagramas aluviais e de *Sankey*, a partir do acompanhamento do progresso de mais de 30.000 alunos de graduação da Universidade de Tecnologia e Economia de Budapeste entre 2010 e 2017. Esta ferramenta permite que os tomadores de decisão tenham uma visão melhorada sobre como os alunos estão progredindo, partindo de uma análise de padrões de fluxo de movimentação. Uma divisão por cores foi usada como auxílio visual nos diagramas (verde representando a graduação e o vermelho significando o abandono) e uma divisão de alunos foi feita por tercis (baseado nos pontos do exame admissional), conforme demonstrado na Figura 3.7. Como resultado destacado pelos autores, a ferramenta tem facilitado a compreensão dos efeitos de algumas mudanças de políticas educacionais nas taxas de retenção e desempenho.

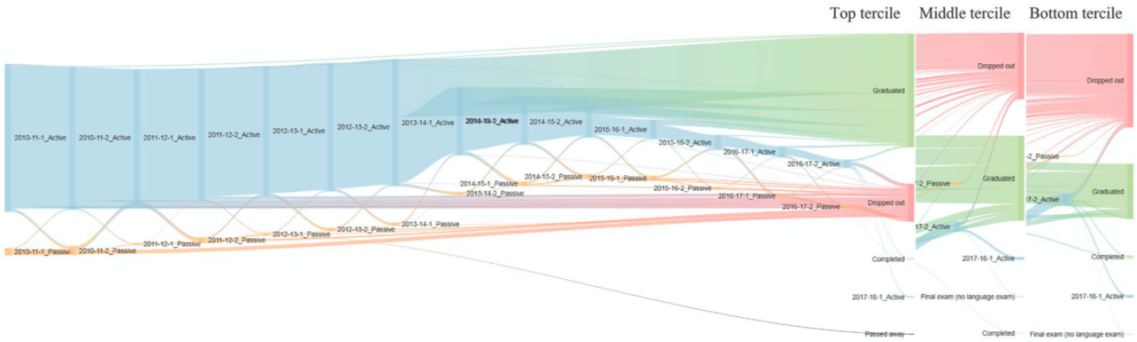


Figura 3.7: Visão de acompanhamento demonstrando a divisão por cores e tercis. Fonte: Horvath et al. [32]

No trabalho de Askinadze et al. [5], são apresentados dados de alunos de um curso de Ciência da Computação em uma Universidade alemã, a partir de um painel educacional. Composto pela combinação de diagramas de *Venn*, *Sankey* e *UpSet*, o objetivo foi permitir uma análise mais detalhada e investigativa dos efeitos dos cursos de forma individual e suas combinações, sob a perspectiva do progresso estudantil. Os diagramas de

Venn e *Upset* foram usados para exibir combinações de exames cumulativamente até um determinado semestre. Para a visualização do progresso de estudantes em diferentes semestres e a partir de diferentes coortes, foram utilizados diagramas de *Sankey*, nos quais, cada nó representa uma combinação de exames em um semestre específico, conforme mostra a Figura 3.8.

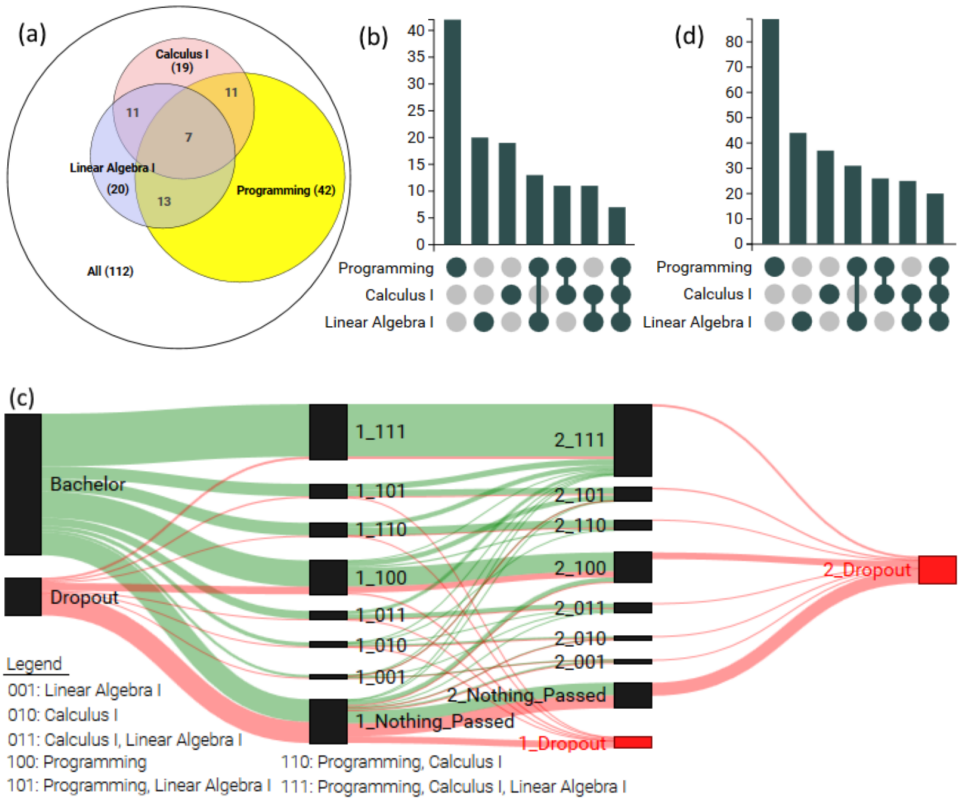


Figura 3.8: Diagramas de *Venn* (a), *UpSet* (b) e *Sankey* (c) mostrando alunos aprovados em exames e que evadiram até o final do segundo semestre. O diagrama de *UpSet* (d) mostrando tentativas de exames. Fonte: Askinadze et al. [5]

A ferramenta de visualização interativa chamada *Graduated Life Explore*, de Gubbala et al. [28], da Universidade do Estado do Sacramento (EUA), foi desenvolvida para que estudantes universitários possam explorar diferentes razões motivacionais para se formar na faculdade em tempo regular. A ferramenta é composta por painéis com filtros diversos (GPA - *Grade Point Average*[8], gênero, grupo étnico, atividade laboral), incluindo visualizações baseadas em diagrama de *Sankey*, combinadas a gráficos de barras e linhas para detalhamentos. Interações por sobreposição de mouse possibilitam a visualização do índice de GPA do aluno e ajudam a compor a ideia central da ferramenta, que é auxiliar os alunos no entendimento dos benefícios de terminar a faculdade em quatro anos, em detrimento das desvantagens de se formar em cinco anos ou mais. A Figura 3.9 apresenta uma visão detalhando o filtro de GPA, no qual cada nó do diagrama representa um agrupa-

[8]GPA é um número que representa o valor médio das notas finais acumuladas obtidas em disciplinas cursadas ao longo do tempo.

mento de alunos de acordo com seu ano de estudos: Primeiro ano (calouro ou *freshman*), Segundo ano (*sophomore*), Terceiro ano (*junior*) e Quarto ano (*senior*). Além disso, relaciona o número de alunos que possuem baixo índice de GPA, que tende a diminuir conforme os alunos chegam ao quarto ano e se formam, enquanto aumenta nos casos em que a graduação ocorre de forma tardia.

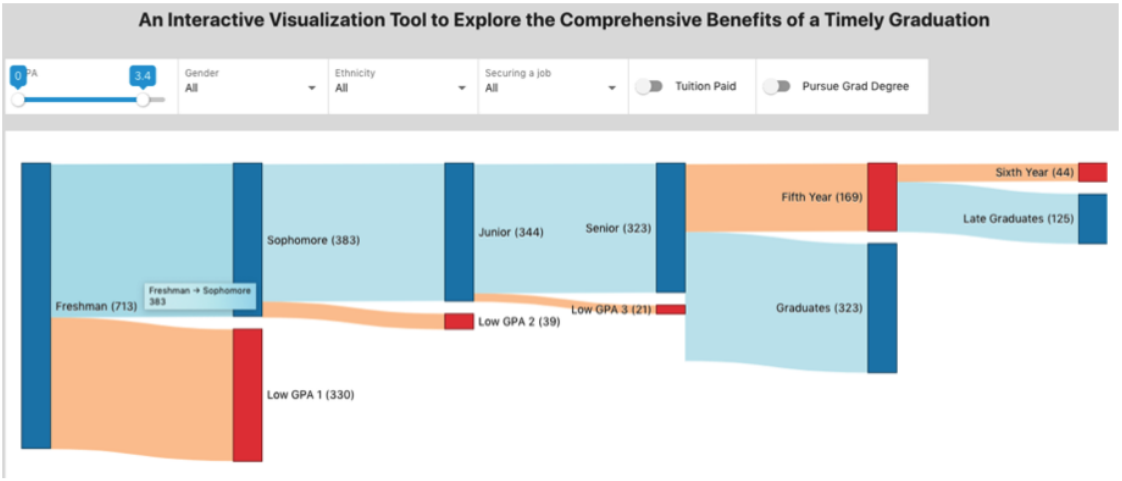


Figura 3.9: Visão detalhando o filtro de (GPA - *Grade Point Average*. Fonte: Gubbala et al. [28]

Baseadas em diagrama de *Sankey*, as visualizações propostas por Klymkowsky et al. [40], são chamadas *Students Progress Visuals* e fornecem informações sobre o progresso e os padrões de percurso dos alunos ao longo do tempo em uma unidade acadêmica da Universidade do Colorado (EUA). Um dos indicadores apresentado como criticos é o tempo de formação dos alunos, citado como preocupante, a partir de um estudo da Academia de Artes e Ciências dos Estados Unidos [4]. O trabalho também discute o processo de transformações dos relatórios tabulares anteriores até a sua evolução para o formato proposto (Figura 3.10), apresentando sua organização de padrão de progresso semestral a partir de quatro grupos de alunos: Matriculado no curso principal original (Azul); Transferido para um curso principal diferente (Verde); Evadido (Vermelho); e Graduado (Roxo).

Na Universidade de Aveiro (POR), Ferreira et al. [24] apresentam uma ferramenta visual chamada FICAvis, como uma evolução do atual sistema de acompanhamento acadêmico. O objetivo deste novo sistema é ajudar a reduzir e prevenir o abandono escolar e aumentar o sucesso acadêmico dos estudantes universitários a partir da visualização de indicadores usando painéis interativos especializados (Acompanhamento Mensal, Comparação Mensal, Performance Geral, Realizações por Indicador/Perfil, Informações Pessoais). Os autores desenvolveram um protótipo de painel usando elementos visuais como gráficos de linhas e barras para demonstrar o progresso de alunos em risco ao longo do tempo a partir do acompanhamento de indicadores como: taxa de aprovação, mensalidade em atraso, aluno bolsista, nota de ingresso, frequência, etc. O trabalho foi desenvolvido utilizando a metodologia de design participativo, detalhando todas as etapas do seu processo

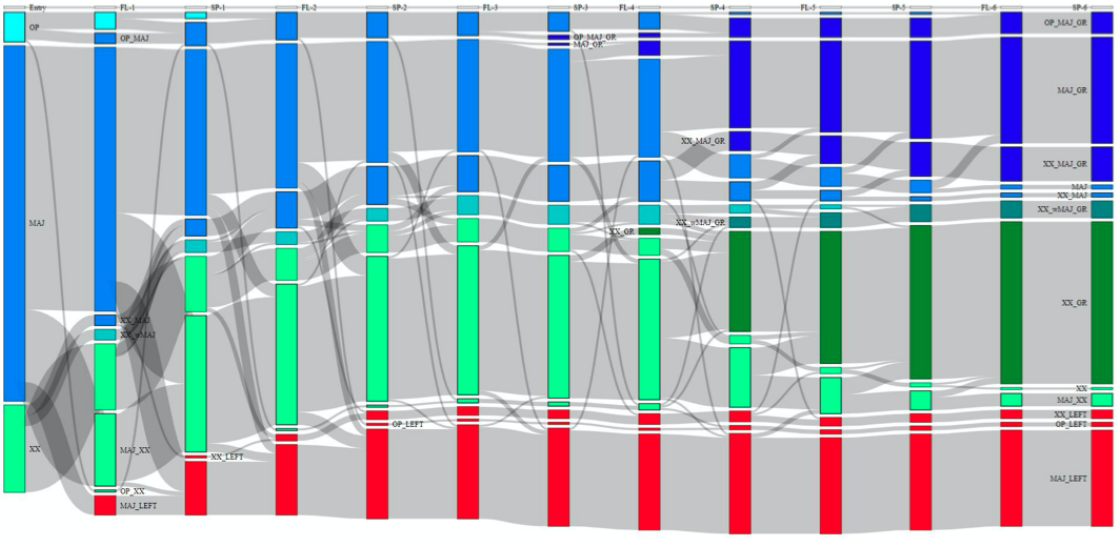


Figura 3.10: Visão do fluxo alunos ao longo dos semestres. Fonte: Klymkowsky et al.[40]

de criação, desde a elicitação de requisitos com base em grupos focais e entrevistas, design e desenvolvimento de protótipo até a sua avaliação. Uma visão do painel de acompanhamento temporal é apresentada na Figura 3.11. O gráfico “Alunos em Risco” (A) é composto por linhas e apresenta, ao longo dos meses, a quantidade de alunos em risco baseado em quatro indicadores: Sucesso - Taxa de sucesso acadêmico abaixo de 50%; Propinas (mensalidade) - Alunos com mensalidades atrasadas; Bolsa - Alunos que solicitaram e não obtiveram uma bolsa de estudos; Assiduidade - Alunos com pelo menos um curso com assiduidade abaixo de 50%. O gráfico “Estado Matrícula” (B) apresenta a distribuição de alunos pelos possíveis estados de matrícula (Ativa, Cancelada, Suspensa - outros motivos, Suspensa - pagamento). O gráfico de comparativo (C) apresenta uma visão detalhada da situação de pagamento dos alunos e o gráfico “Anulações Acumuladas” (D) apresenta uma visão mensal de cancelamentos e seu valor acumulado.

O trabalho de Skurla et al. [68], da Universidade de Baylor (Texas - EUA), apresenta uma ferramenta visual em que a análise do movimento dos alunos pelos currículos de Engenharia foi tratada como um problema de fluxo. Utilizando um aplicativo originalmente usado para análise de fluxos de materiais (e!Sankey), visualizações foram criadas a partir de diferentes agrupamentos, empregando padrões de fluxo presentes nos diagramas de *Sankey*. Na visualização, setas representam o fluxo de alunos, dimensionadas em relação ao número original de indivíduos em uma coorte específico, que dão uma ideia rápida para onde esses alunos estão se direcionando: para a graduação (com um diploma de engenharia), para a retenção (dentro da universidade em outro curso) ou para a saída da universidade (evasão), como demonstra a Figura 3.12. Os autores mencionam que o método gráfico aplicado ao fluxo de alunos e a capacidade de detalhamento (*drill-down*), permitiram a rápida compreensão de uma grande quantidade de dados e tornaram-se

Figura 3.11: Visão da interface de acompanhamento temporal de alunos. Fonte: Ferreira et al. [24]

ferramentas valiosas para a avaliação das medidas de gestão de retenção e matrículas anuais.

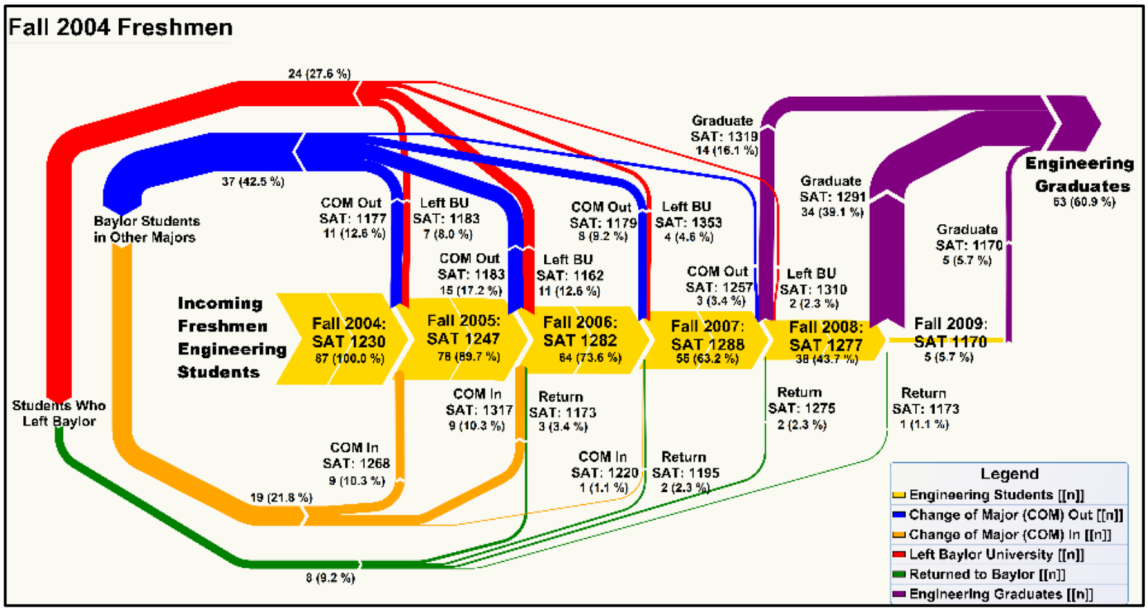


Figura 3.12: Visão do fluxo de entrada de novos alunos e progresso ao longo de 5 anos. Fonte: Skurla et al. [68]

O'Handley et al. [50], apresentam a ferramenta *CoursePathVis*, que utiliza diagrama de *Sankey* como base para uma análise visual do progresso dos alunos em um

currículo universitário (Universidade de Notre Dame - EUA). Nesta ferramenta, alunos de quatro coortes (de 2019 a 2022) em um departamento são agrupados de várias maneiras: por seus cursos de AP[9], cursos semestrais e um curso usado como funil (a ser especificado pelo usuário) para uma melhor entendimento sobre como progridem ao longo do tempo. A funcionalidade de “curso de funil” é descrita como uma forma de agrupamento que auxilia no entendimento a respeito de como um grupo de cursos (agora como uma única entidade) se conecta e afeta a realização de um curso subsequente, ou como os alunos progridem até o curso de funil em questão, conforme pode ser visto na Figura 3.13.

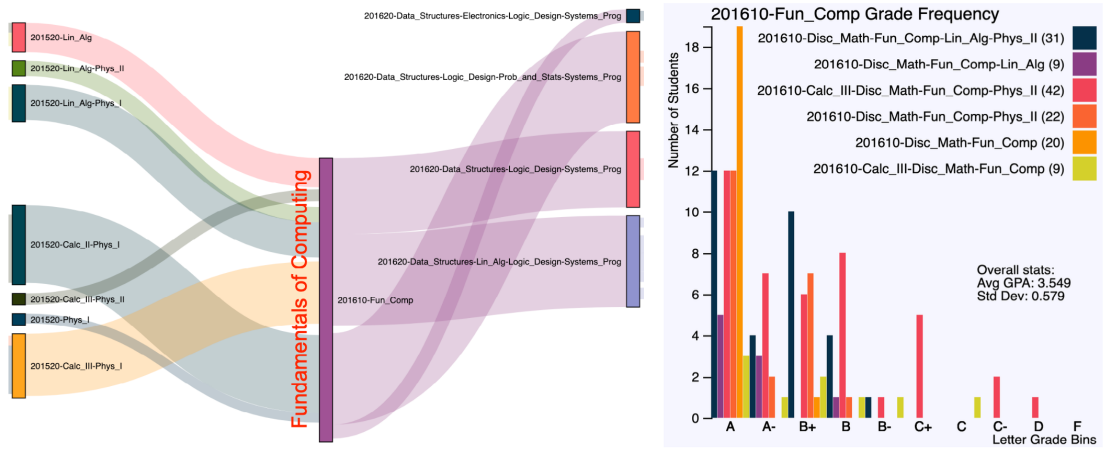


Figura 3.13: Destaque da visão funil da movimentação entre as disciplinas do curso de Fundamentos da Computação. Histograma detalhado do curso funil e suas disciplinas (*term-course*). Fonte: O’Handley et al. [50]

3.2 Análise dos Trabalhos

A Tabela 3.2, sintetiza o estudo da RSL apresentando as diferenças e semelhanças entre os trabalhos (ordenados por data de publicação) no que diz respeito a técnicas de visualização, interações utilizadas e técnicas estatísticas. Nesta tabela é possível perceber que a técnica de visualização mais utilizada pelos trabalhos é o diagrama de *Sankey*, seja de forma exclusiva ou complementada por outro tipo de gráfico.

A partir da análise detalhada dos trabalhos e seguindo os direcionamentos apresentados pelas perguntas norteadoras da RSL, foi possível respondê-las, conforme apresentado a seguir:

QP1 - Como são visualizados os dados de fluxo e progresso dos alunos?

Considerando que todos os trabalhos selecionados possibilitam que um fluxo de alunos seja acompanhado ao longo de um período de tempo, uma primeira observação

[9]Advanced Placement - Programa (EUA, CAN) no qual estudantes de Ensino Médio podem frequentar disciplinas de uma instituição de ensino superior e receber uma avaliação por seu desempenho.

Tabela 3.2: Comparação entre os trabalhos analisados na RSL.

Categoria	Funcionalidade	Heilemann et al. 2015 [31]	Greer et al. 2016 [27]	Raji et al. 2017 [53]	Basavaraj et al. 2018 [8]	Vaclavek et al. 2018 [72]	Horváth et al. 2018 [32]	Askinadze et al. 2019 [5]	Gubbala et al. 2019 [28]	Klymkowsky et al. 2019 [40]	Ferreira et al. 2020 [24]	Skurla et al. 2021 [68]	O'Handley et al. 2022 [50]
Técnicas de Visualização	Árvore radial			✓									
	Gráfico de pizza			✓							✓		
	Diagrama de Sankey	✓	✓	✓	✓		✓	✓	✓	✓		✓	✓
	Histogramas					✓							✓
	Gráficos de coluna / barras / linhas				✓	✓			✓		✓		
	Diagrama de Venn e gráfico UpSet							✓					
	Tabelas										✓		
Tipos de Interação	Navegação e destaque de fluxos utilizando mouse.	✓	✓	✓	✓		✓			✓			
	Filtros na aplicação	✓	✓			✓			✓		✓		✓
	Interação por mouse.	✓	✓			✓			✓		✓		✓
	Filtros antes de carregar a aplicação						✓						
	Navegação em detalhamento (*drill-down*)	✓	✓									✓	
Técnica Estatística	Correlação entre cursos, utilizando coeficiente de correlação de Pearson.			✓	✓								
	Teorema de Bayes para agrupamentos de alunos em grupos de acompanhamento.					✓							
	Agrupamento de alunos por nota de ingresso, usando tercis.						✓						

é que diagramas de *Sankey* aparecem sendo utilizados com sucesso como visualização principal [31, 27, 8, 32, 28, 40, 68, 50], como visualização complementar [5], ou ainda adaptado, como uma variação [53].

Outras formas populares de visualização para sequências de dados, como gráfico de coordenadas paralelas e gráfico de área empilhada foram avaliadas por O'Handley et al. [50] e se mostraram pouco apropriadas por não contemplarem completamente as necessidades de visão agrupada, agregada e de fluxos de alunos ao longo de um período de tempo. A Figura 3.14 apresenta a comparação entre estes tipos de gráficos: (A) Diagrama de *Sankey*, que contempla todas as necessidades de agrupamento, agregação e visão por período; (B) Coordenadas Paralelas, que não suportam uma visão agregada; e (C) gráfico de Área Empilhada, que não permite a visão de agrupamentos e múltiplos fluxo simultaneamente.

Corroborando com esta observação, um estudo feito por Klymkowsky et al. [40], apresentado na figura 3.15, descreve o esforço de tentar acompanhar as mudanças no

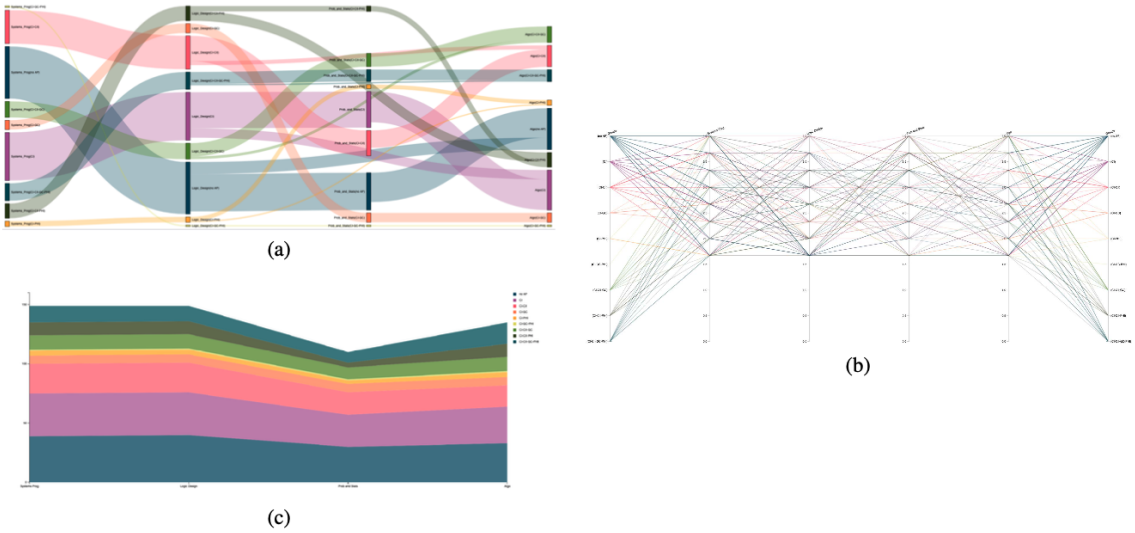


Figura 3.14: Estudo comparando formas de visualização de dados sequenciais: (A) Diagrama de *Sankey*; (B) Coordenadas Paralelas; (C) gráfico de Área Empilhada. Fonte: O'Handley et al. [50]

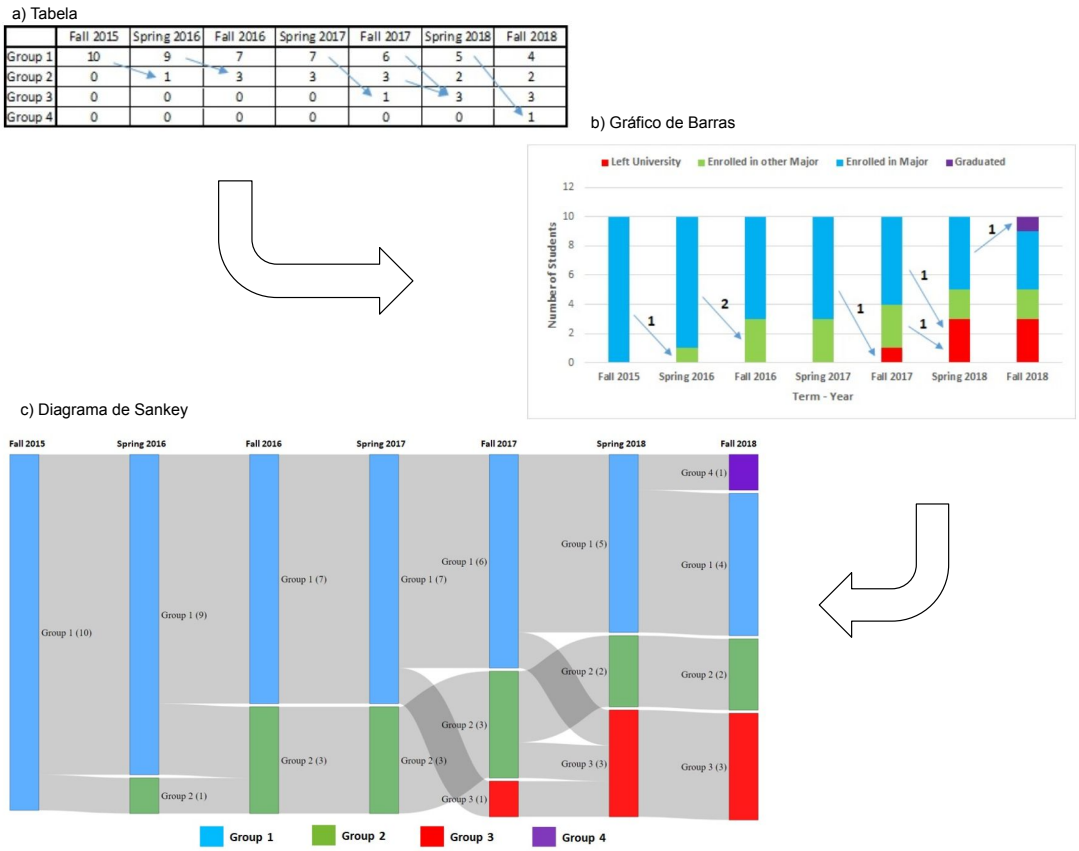
número de alunos entre os diferentes grupos ao longo de vários semestres usando tabelas ou gráficos de barras. Ainda conforme os autores, o uso de tabelas ou gráficos de barra poderia ser uma barreira, principalmente para pessoas não familiarizadas com os dados dos alunos, sugerindo a utilização do diagrama de *Sankey* como uma forma eficiente de visualização.

Ainda, nos trabalhos avaliados, outras técnicas complementares de visualização são utilizadas: como Gráficos de Barras [72, 28, 24], Gráficos de Linhas [8, 28, 24] e Gráficos de Pizza [53, 24]. Histogramas são usados em alguns casos para ajudar a compor as visões de detalhamento [72, 50]. Entre as técnicas de interação disponibilizadas, navegação por mouse, realce de informações e filtros são as mais comuns e disponíveis em quase todos os trabalhos [72, 24, 50]. Horvath et al. [32] apresenta uma variação, propondo a utilização de filtros aplicados antes da visualização estar disponível, devido a grande quantidade de dados envolvidos.

QP2 - Por meio das visualizações apresentadas, quais resultados comuns foram encontradas na literatura, objetivando promover uma melhoria no cenário educacional?

Complementando as análises sobre as visualizações propriamente ditas (QP1), outras características compõem importantes aspectos sob os quais os trabalhos foram desenvolvidos, como o conjunto de dados estudado, seu público alvo, se fizeram ou não uso de técnicas estatísticas para auxiliar no entendimento e predizer evasões, e, claro, seus objetivos e resultados em prol da melhoria no cenário educacional (QP2).

Os conjuntos de dados utilizados nos trabalhos alternam entre: dados de toda a instituição [53, 24], restritos aos cursos em seus departamentos [8, 68] ou ainda de toda



	Fall 2015	Spring 2016	Fall 2016	Spring 2017	Fall 2017	Spring 2018	Fall 2018
Group 1	10	9	7	7	6	5	4
Group 2	0	1	3	3	3	2	2
Group 3	0	0	0	0	1	3	3
Group 4	0	0	0	0	0	0	1

Figura 3.15: Estudo realizado e que demonstra a evolução de uma visualização de dados estudantis a partir de uma tabela (A), passando por um gráfico de barras (B), até sua forma final detalhada em um diagrama de *Sankey* (C). Fonte: Klymkowsky et al. [40]

uma faculdade em específico [72]. Um grande obstáculo que as instituições enfrentam é a dificuldade de gerenciar, manipular e processar estes grandes volumes de dados (*big data*), tornando difícil determinar até mesmo por onde iniciar as análises e podendo levar os pesquisadores a restringirem seu escopo inicial [31].

O público alvo dos trabalhos são agentes das instituições de ensino como gestores, técnicos e educadores interessados na melhoria dos indicadores educacionais, com exceção do trabalho de Gubbala et al. [28] que apresenta um sistema no qual o estudante é instigado a realizar o seu próprio gerenciamento de percurso e perceber possíveis pontos de melhoria.

Um importante ponto foi observado a respeito dos trabalhos em relação a utilização de métodos estatísticos. A utilização ou não destes métodos pode ser considerado uma espécie de “divisor de águas”, como proposto por Klymkowsky et al. [40], que sugerem o agrupamento dos trabalhos da área em duas categorias de acordo com seus objetivos: Descritivos ou Preditivos. A primeira tem como base somente o uso de “Métodos Descritivos”, pretendendo descrever (resumir) com precisão o desempenho acadêmico dos alunos, por meio de apresentação de gráficos e tabelas, e delegando aos usuários destas ferramentas a maior parte do trabalho de análise. Em outra categoria podemos

incluir todos os trabalhos que além de descreverem o progresso de estudantes, procuram utilizar métodos estatísticos para predizer o desempenho dos alunos.

Nos trabalhos avaliados são mencionados alguns métodos estatísticos não preditivos [72, 32], que usam agrupamentos para auxiliar na identificação de padrões de desempenho, enquanto Raji et al.[53] e Basavaraj et al.[8], buscam apresentar a correlação entre os cursos. Sobre modelos preditivos, a literatura [19, 37, 2] tem apresentado resultados positivos, nos quais os mesmos são usados com sucesso para identificar alunos em risco e auxiliar tomadores de decisão a intervir proativamente, embora Charleer et al. [15] mencionem que esses modelos preditivos de evasão podem ser caixas pretas, pois não fornecem aos usuários informações sobre os motivos das decisões tomadas (em relação à evasão). Esta lacuna encontrada nos trabalhos analisados, aparentemente representa uma oportunidade de estudos, através da qual a utilização de uma visualização especializada de percurso estudantil pode ser conectada a um conjunto de métodos estatísticos preditivos. Esta descoberta corrobora com os comentários de Raji et al., Vaclavek et al. e Greer et al. [53, 72, 27] que mencionam a possibilidade de implementar futuramente este tipo de modelo nas suas visualizações.

Sobre os objetivos e resultados encontrados, o acompanhamento de alunos e seu percurso até o sucesso ou evasão é o direcionador principal de trabalhos [8, 68, 24] que afirmam terem alcançado resultados positivos a partir das análises feitas por responsáveis entrevistados utilizando suas visualizações propostas. A verificação e acompanhamento de políticas educacionais foi um tema explorado em outro trabalho [32], juntamente com a relação (comparação) entre cursos e seus efeitos [8, 5]. O acompanhamento de tempo de formação (tempo de permanência no curso) [31, 28, 40], o acompanhamento de padrões de mobilidade estudantil [31, 40] e o aperfeiçoamento do planejamento de matrículas anuais [68], foram outros temas citados.

Concluindo as análises acerca da literatura, dentre as limitações identificadas, além da já citada sobre a utilização de técnicas estatísticas, estão os desafios de explorar visualmente um grande volume de dados de forma temporal, unificada e completa, além de permitir uma visão mais individual e que vá ao encontro das necessidades dos educadores e administradores. Esta questão também foi apontada por Horvath et al. [32]. Por fim, alguns trabalhos analisados não deixam claro se as propostas podem ser utilizadas com dados de outras instituições. As exceções são Horvath et al. [32], que pretendem contemplar este quesito em trabalhos futuros e Greer et al. [27], que mencionam ter utilizado dados do sistema TEA (*Tools for Evidence-Based Action*) integrado a outras universidades.

4. PROTÓTIPO DE VISUALIZAÇÃO DO PERCURSO ACADÊMICO

Considerando as oportunidades e limitações encontradas através da análise dos trabalhos resultantes da RSL, foi implementado um protótipo baseado no uso de diagramas de *Sankey* para auxiliar a percepção do fluxo de alunos entre as disciplinas em um contexto temporal. Segundo Heileman et al. [31] e Riehmann et al. [57], os mesmos são apropriados para a visualização e entendimento de dados complexos como os abordados neste trabalho. Neste capítulo são apresentados detalhes sobre o conjunto de dados usados no protótipo (Seção 4.1), o ambiente de implementação utilizado (Seção 4.2) e a visualização criada usando diagramas de *Sankey* (Seção 4.3).

4.1 Conjunto de Dados

Para este trabalho foi usado um conjunto de dados brutos, recebidos em formato CSV (*Comma-Separated Values*), de uma instituição de ensino superior privada, apresentando cerca de 4500 registros de 530 alunos de 3 cursos e abrangendo o período de 2013/1 até 2019/1.

As dimensões disponíveis, originalmente, neste conjunto de dados são apresentadas no dicionário de dados da Tabela 4.1.

Dimensão	**Tipo**	**Descrição**
curso	TEXTO	Nome do curso (ex: Curso 1)
formadeingresso	TEXTO	Forma de ingresso (ex: Vestibular)
formadeegresso	TEXTO	Forma de egresso (ex: Formado)
semestredeingresso	TEXTO	Primeiro semestre na instituição (Ex: 2019/1)
ultimosemestrematriculado	TEXTO	Último semestre na instituição (Ex: 2019/1)
situacaonadisciplina	TEXTO	Situação no final da disciplina (Ex: Aprovado)
semestredadisciplina	TEXTO	Semestre da disciplina (Ex: 2019/1)
codigodapessoa	INTEIRO	Identificador do aluno (Ex: 1234)
codigodadisciplina	TEXTO	Identificador da disciplina (Ex: Disciplina 1234)

Tabela 4.1: Dicionário de dados original do conjunto.

Todos os dados utilizados foram previamente anonimizados para preservar a identidade dos alunos envolvidos e fornecidos mediante solicitação junto à direção da instituição para atenderem os requisitos de privacidade exigidos pela Lei Geral de Proteção de Dados. Por este motivo, de preservar a identidade dos estudantes, não foram incluídos dados socioeconômicos dos mesmos. O conjunto de dados usado está restrito a este estudo, não sendo possível sua divulgação externa.

A Lei nº13.709, de 14 de agosto de 2018, conhecida como Lei Geral de Proteção de Dados (LGPD), consiste em uma norma legal que objetiva a garantia da segurança de dados pessoais, além de trazer importantes alterações na Lei nº 12.965/2014 ou Marco Civil da Internet [11]. A LGPD, em relação a pesquisa acadêmica, apresenta algumas condições dispostas nos artigos, segundo Hartmann [30]:

- Art. 4º Esta Lei não se aplica ao tratamento de dados pessoais:
  - II - realizado para fins exclusivamente: b) acadêmicos, aplicando-se a esta hipótese os arts. 7º e 11 desta Lei;
- Art. 7º O tratamento de dados pessoais somente poderá ser realizado nas seguintes hipóteses:
  - IV - para a realização de estudos por órgão de pesquisa, garantida, sempre que possível, a anonimização dos dados pessoais;
- Art. 11. O tratamento de dados pessoais sensíveis somente poderá ocorrer nas seguintes hipóteses:
  - II - sem fornecimento de consentimento do titular, nas hipóteses em que for indispensável para:
    a) Fins Exclusivamente Acadêmicos (art. 4, II, b);
    b) Para Realização de Estudos (arts. 7, IV e 11, II, c);
    c) Por Órgão de Pesquisa (arts. 5, XVIII, 7, IV e 11, II, c);
    d) Obrigação de Anonimização: art. 5o, XI;
    e) Obrigação de Finalidade, Boa-fé, Interesse Público (art. 7, § 3º).

4.2 Implementação

Para a implementação do protótipo e melhor manipulação, mapeamento e criação de filtros, os dados brutos foram passados para uma base de dados relacional *SQLite*[1], que permite o uso do padrão SQL para consulta e manipulação de dados de maneira mais avançada do que o uso direto de arquivos [49]. Uma API (*Application Programming Interface*, ou Interface de Programação de Aplicativos) de consulta foi criada para a integração desta base de dados com a camada de visualização que é acessada pelo usuário através de um navegador web. Outra função desta API é criar a organização e agrupamento dos dados necessários para geração da estrutura de DAG[2] (*Directed Acyclic Graph*, ou grafos

[1] *SQLite*: banco de dados embarcado, de código aberto.
[2] DAG - *Directed Acyclic Graph*: grafos sem ciclos e com ligações conectadas em uma mesma direção [46].

acíclico dirigidos) [46], utilizada pelo diagrama de *Sankey* para construção da visualização, além de possibilitar a implementação de filtros que podem ser aplicados a partir da interface de usuário. O protótipo foi implementado na linguagem PHP[3], apropriada para o desenvolvimento web [7].

Para a criação do diagrama de *Sankey* no protótipo, foi utilizada, inicialmente, a biblioteca gráfica D3.js, baseada na tecnologia SVG *(Scalable Vector Graphics)* e comumente utilizada em visualizações interativas [10]. Após um teste inicial notou-se que a performance oferecida pela biblioteca não era a ideal, corroborando com o estudo de desempenho feito por Kee et al. [38]. Portanto, devido ao grande número de nós e ligações presentes na visão integral do conjunto de dados e como forma de tornar a performance mais fluida na exploração da visualização, optou-se pela utilização da biblioteca *Echarts* [43] para a renderização do diagrama *Sankey*, que é baseada na tecnologia de *Canvas* e incluída na maioria dos navegadores [47]. Outra necessidade encontrada durante a implementação foi a respeito da navegação pelo diagrama. Como a biblioteca utilizada não possuía nativamente as possibilidade de navegação panorâmica e *zoom*, estas foram implementadas utilizando como base a projeto de código aberto chamado de *panzoom* [4].

4.3 Visualização

A proposta de visualização foi desenvolvida a partir da ideia do *pipeline* proposto por Card [14], que apresenta as etapas que devem ser seguidas para chegar a uma representação visual a partir dos dados. Assim, a etapa inicial consiste na obtenção e conversão dos dados de entrada brutos, seguida do agrupamento de informações. Na sequência, foram definidos os processos de mapeamento e filtragem de dados até chegar à visualização, resultando no fluxo de visualização proposto e apresentado na Figura 4.1.

O ponto de partida do protótipo de visualização é um dos cursos disponíveis. Filtrados através da interface, cada curso selecionado permite a visualização de um conjunto de semestres disponíveis.

A visualização usando diagrama de *Sankey* foi implementada inspirada nos trabalhos relacionados [31, 53, 32, 28] e tem como principal vantagem o uso de nós e ligações com espessura variável, proporcionais à magnitude dos fluxos em questão [59]. No protótipo esta característica foi empregada no acompanhamento da quantidade de alunos em cada semestre, disciplina e demais agrupamentos.

A Figura 4.2 apresenta a visão geral de um curso no diagrama de *Sankey* do protótipo. O detalhe destacado mostra o visão criada para a análise de dados de um semestre, a partir dos fluxos de alunos originários de um agrupamento semestral a esquerda,

[3] https://www.php.net/
[4] https://github.com/anvaka/panzoom

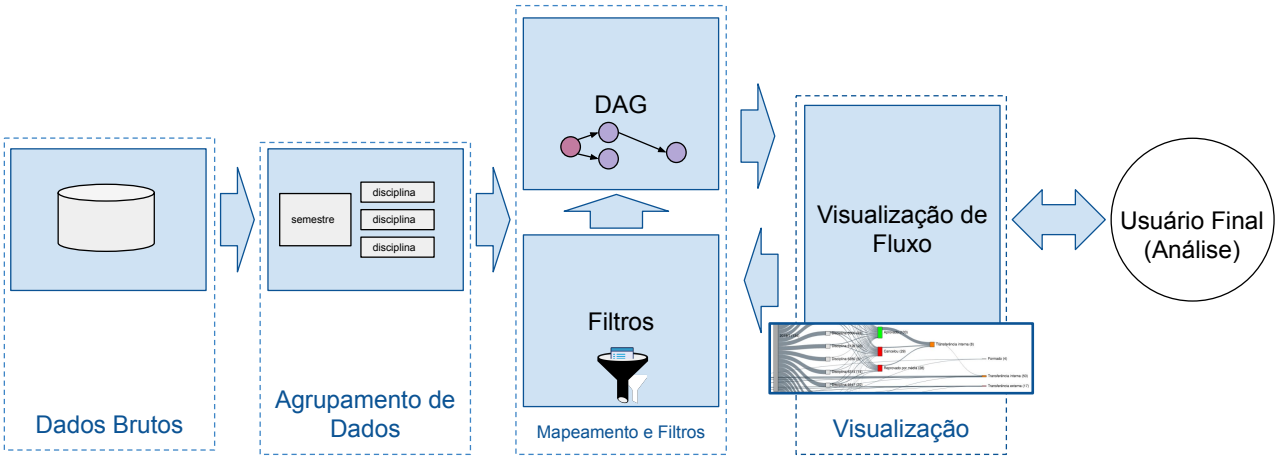


Figura 4.1: *Pipeline* de visualização proposto.

passando por suas disciplinas em direção aos estados de Aprovado, Cancelou, Reprovado por média, Transferência Interna, Transferência Externa e Formado.

Figura 4.2: Visão geral de um curso a partir do protótipo de Visualização do Percurso Acadêmico.

Os dados foram organizados em dois blocos de nós conectados: Agrupamento de Semestre (Semestre - Ingressos e Forma de Egressos) e Agrupamento de Disciplinas (Disciplina e Situação na Disciplina). Estes nós foram dispostos ao longo de um período de tempo em um grande arranjo de forma a representarem um semestre, em um curso na instituição.

Esta organização foi criada a partir do modelo visual proposto por Horváth et al. [32], adaptada para um melhor entendimento da relação entre as diferentes movimentações envolvendo disciplinas e semestres, em conjunto com os agrupamentos realizados nas diferentes dimensões de dados estudantis.

A Figura 4.3 mostra a representação do modelo de dados proposto para o protótipo e a sua aplicação a partir da visão de um semestre, na qual a organização é destacada para exemplificar a composição. Adicionalmente, cada nó possui um rótulo de nome e o número correspondente de alunos únicos associados a ele. Portanto, o bloco de “Semestre” se conecta em múltiplos pontos aos blocos de “Disciplina” (alunos matriculados em uma ou mais cadeiras), que conecta com um ou mais blocos de “Situação na Disciplina”, que por sua vez, se conecta ao próximo bloco de “Semestre” e de “Forma de Egresso”.

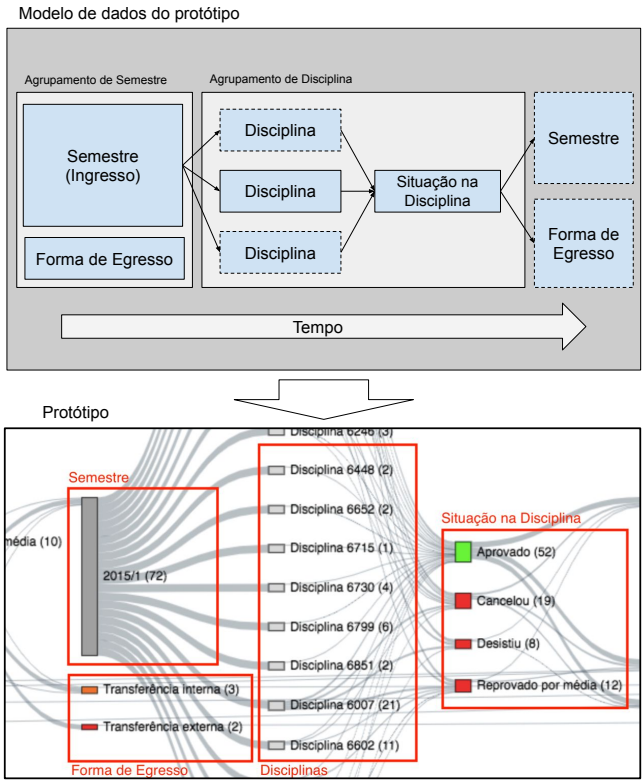


Figura 4.3: Representação do modelo de dados e a sua aplicação no protótipo, com destaque na organização de nós.

Grupos baseados na “situacaonadisciplina” e na “formadeegresso” foram criados como forma de entender a proporção de alunos em situação positiva (Aprovado), e nas demais condições (Cancelou, Desistiu, Reprovado por média, Reprovado por faltas, Transferência Interna, Transferência Externa). Desta maneira, para cada agrupamento de “semestredadisciplina”, “codigodadisciplina”, fluxos com as condições descritas foram criados baseados nos indivíduos com tal resultado, conforme observado na Figura 4.4. Nesta figura, é possível observar o bloco referente ao semestre 2017/2, o fluxo em destaque, obtido pela interação do usuário com o mouse sobre o nó “Disciplina 6767”, e o seu respectivo rótulo, contendo uma listagem dos identificadores de alunos que cursaram a disciplina (7).

De acordo com Lin et al. [44], também como forma de facilitar a exploração, foram utilizadas cores significantes (semanticamente ressonantes) para representar os agrupamentos. Por exemplo, para o agrupamento de “situacaonadisciplina” foi utilizada

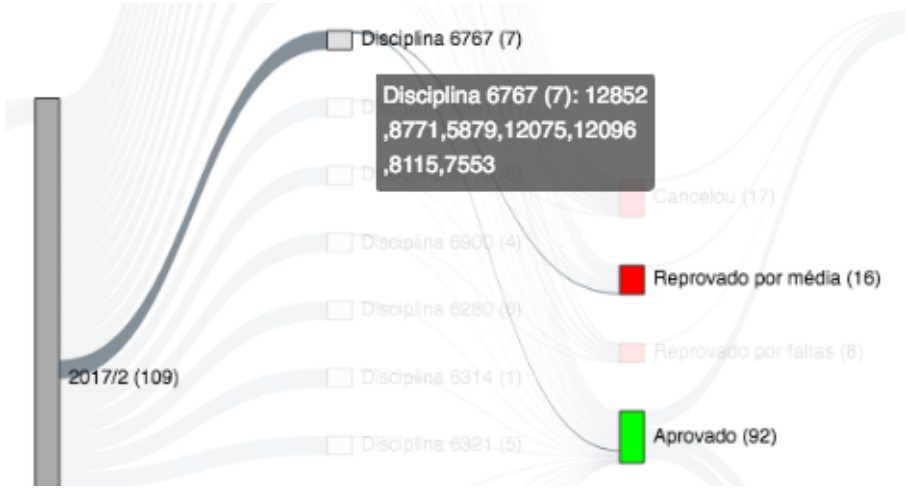


Figura 4.4: Visão ampliada de um bloco de semestre, destacando o fluxo individual de uma disciplina.

a cor verde para representar a situação positiva de “Aprovação” e a vermelha para “cancelou” e demais situações negativas. No caso do agrupamento “formadeegresso”, a cor laranja foi utilizada para a situação de “Transferência Interna”, representando a necessidade de uma maior atenção e a vermelha para “Transferência Externa”. A cor cinza foi utilizada de forma geral para reduzir a atenção visual nos demais agrupamentos. A lista completa das representações de cores é apresentada na Tabela 4.2.

Tabela 4.2: Referência de cores utilizadas para os agrupamentos.

Situação na Disciplina (situacaonadisciplina)	Forma de Egresso (formadeegresso)	Cor
Aprovado	Formado	**Verde**
[não usado]	Transferência Interna	**Laranja**
Cancelou, Desistiu, Reprovado por média, Reprovado por faltas	Transferência Externa	**Vermelha**

Para os agrupamentos derivados de cada disciplina, somente indivíduos únicos foram contabilizados. Desta forma o montante de alunos de um semestre nunca será igual a soma dos alunos em cada disciplina. Uma situação comum é quando um mesmo indivíduo cursa múltiplas disciplinas, é aprovado nestas, porém no agrupamento de “Aprovado” é contabilizado apenas uma vez. Este comportamento, que permite a inclusão de um único indivíduo por agrupamento, possibilita com que o usuário aplique um filtro por estudante e assim consiga observar seu percurso, individualmente, ao longo do período acadêmico.

Para facilitar a exploração de dados, o protótipo apresenta filtros de seleção para um dos três cursos separadamente e por um ou mais alunos de forma simultânea. As interações disponíveis são: de movimentação panorâmica, ampliação e redução da visualização, detalhamento (*tooltip*) de alunos em cada nó/ligação a partir do movimento

“pairar sobre” / *hover*. Também utilizando este movimento é possível perceber um destaque visual no qual os nós e ligações conectados permanecem na suas cores definidas, enquanto os demais elementos da visualização são esmaecidos com a intenção de reduzir seu destaque como na Figura 4.4. O filtro de alunos é automaticamente populado e a visualização é atualizada quando um nó/ligação é clicado, tornando simples um visão destacada do percurso para um ou mais alunos, conforme ilustra a Figura 4.5.

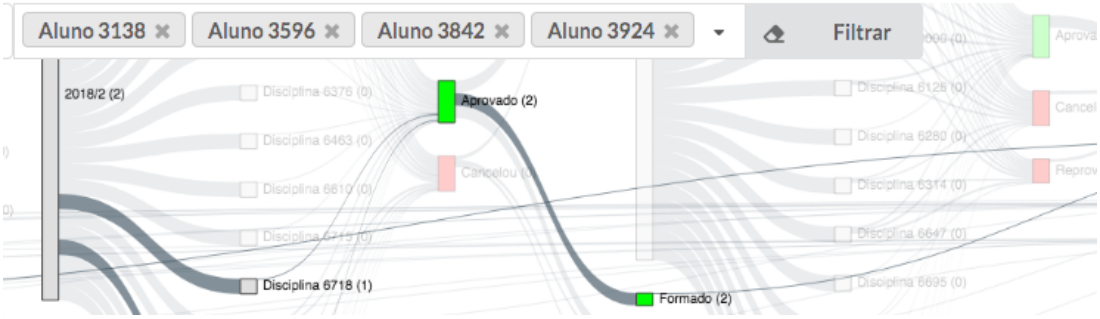


Figura 4.5: Visão ampliada da interface de filtro e destaque do progresso por aluno(s).

Este protótipo preliminar foi apresentado para especialistas de domínio durante as entrevistas descritas no Capítulo 5, incentivando-os a fazerem uma análise exploratória dos dados fornecidos [35].

5. ENTREVISTAS COM ESPECIALISTAS DE DOMÍNIO

Visando auxiliar na identificação de requisitos para este estudo e das necessidades e ferramentas utilizadas, foram feitas entrevistas semiestruturadas em profundidade com especialistas de domínio. Este tipo de entrevista é mais interativa, os pesquisadores podem esclarecer perguntas dos participantes e investigar respostas inesperadas, usando um roteiro com perguntas abertas e uma condução mais informal [45, 66, 60].

O processo de entrevistas e perfil dos participantes são detalhados na Seção 5.1. Uma análise e discussão das entrevistas é feita na Seção 5.2. Os resultados, incluindo a lista de requisitos identificada, são apresentados na Seção 5.3

5.1 Processo de Entrevista e Perfil dos Participantes

As entrevistas foram guiadas por um questionário semiestruturado cujas questões foram divididas em quatro seções, levando a um total de 27 perguntas abertas. Estas seções são: (1) levantamento de perfil; (2) práticas e ferramentas, perfil do estudante (disciplinas e mudanças de curso) e necessidades dos especialistas; (3) análise do protótipo de visualização; (4) considerações finais. O objetivo foi ampliar o conhecimento do cenário atual de ferramentas e métodos usados pelas instituições de ensino para analisar o percurso dos estudantes e também apresentar o protótipo implementado para receber comentários, contribuindo com o processo de elicitação. O protocolo seguido para estas entrevistas foi aprovado pelo Comitê de Ética em Pesquisa (registro CAAE 29383420.0.0000.5336) e o questionário aplicado se encontra no Anexo A juntamente com o modelo de TCLE (Termo de Consentimento Livre e Esclarecido), disponível no Anexo C. Todas as entrevistas ocorreram remotamente e duraram de 40 a 60 minutos.

Quatro profissionais foram entrevistados (três homens e uma mulher), de três instituições (duas privadas e uma pública), das áreas da Ciência da Computação e Administração. Quanto a formação, três participantes eram pós-graduados (mestrado e doutorado) e um tinha somente o curso de bacharelado. A experiência deles variou de 5 à 10 anos na atividade de gestão ou acompanhamento de alunos no ensino superior, de forma direta ou participando de grupos de trabalho. A Figura 5.1, sintetiza os dados sobre o perfil dos participantes.

O número reduzido de participantes se deve ao fato que, a entrevista em profundidade é uma técnica de pesquisa qualitativa, na qual um pequeno número de pessoas são envolvidas de maneira mais intensiva e individual, de forma a explorar suas perspectivas sobre uma determinada ideia, programa ou situação [48]. Corroborando com a afirmação anterior, Lazar et al [42], entendem que, para o levantamento requisitos com

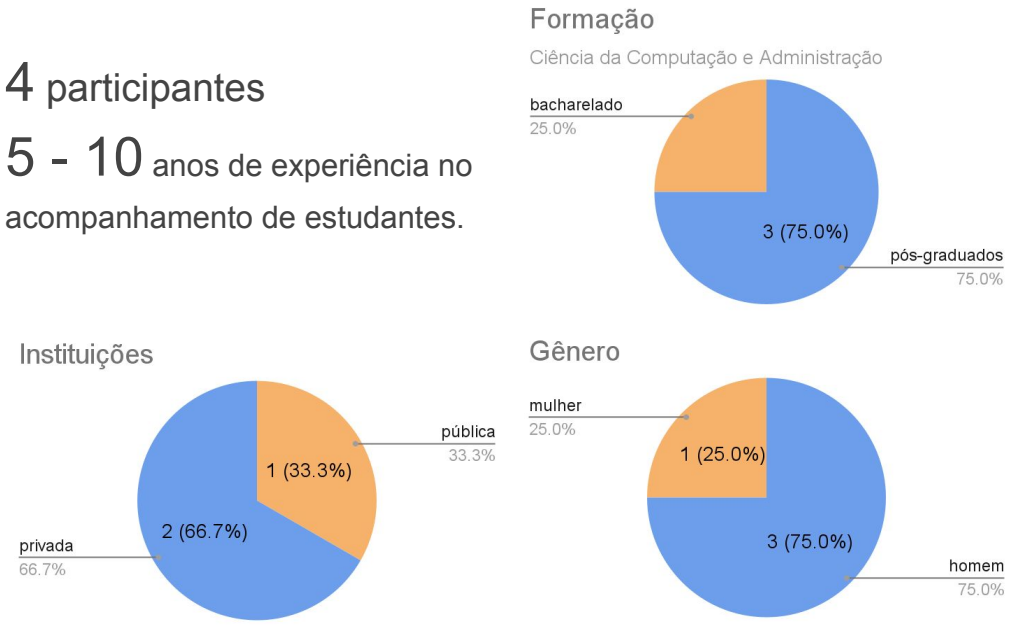


Figura 5.1: Perfil dos participantes.

especialistas de domínio, discussões aprofundadas com dois ou três indivíduos motivados podem fornecer dados suficientes.

5.2 Análise e Discussão das Entrevistas

A seguir é apresentada uma análise das entrevistas realizadas e sua discussão. As análises das entrevistas foram realizadas a partir da transcrição dos áudios, gravados com o consentimento dos participantes. Devido ao dinamismo e informalidade da entrevistas, as perguntas usadas para guiar o trabalho, não foram feitas em uma mesma ordem (para todos os participantes), o que levou a um agrupamento, para discussão, menos formal, mas não menos rico, e que é apresentado nesta seção. No corpo do texto, também, são referenciados os requisitos levantados e que são apresentados de forma detalhada ao final da seção.

Uma primeira observação a partir dos especialistas ouvidos é que todas as instituições de ensino possuem alguma ferramenta para o acompanhamento dos seus estudantes. O acompanhamento de grupos é feito (R3), mesmo que apenas semestralmente, para obter uma visão geral do período e fornecer subsídios para ações de melhoria futuras (abertura de novas turmas, realocação de professores, indicadores de qualidade da instituição, etc). Porém, o nível destas análises e o ferramental utilizado varia bastante entre elas. Um dos profissionais comentou sobre a periodicidade de indicadores dentro de um semestre: *"Existem indicadores que acontecem no meio do semestre e têm indicadores que acontecem no final do semestre"*. Um participante comentou que o acesso aos dados

institucionais de uma forma mais ampla é restrito, requerendo uma série de autorizações que inviabilizam novas iniciativas: *"Tenho uma certa decepção por a gente não ter os dados brutos dos alunos"*.

Considerando as ferramentas utilizadas, todas as instituições usam múltiplas soluções desenvolvidas internamente para acompanhamento e análise de dados estudantis. Um participante mencionou que: *"... temos essa ferramenta [para análise de indicadores usada a nível departamental] que queremos evoluir para que vire uma ferramenta para a universidade"*. Planilhas eletrônicas também são utilizadas por todas as instituições, embora apontadas como limitadas para o acompanhamento de um grande volume de dados. O uso da linguagem R[1] foi mencionada por um participante, responsável pela instituição que possui um ferramental mais "maduros" (incluindo softwares de mercado), para avaliação de dados estudantis, utilizando indicadores em tempo real e filtros para análises mais aprofundadas (R11). Nesta instituição, foram entrevistados dois profissionais, que avaliavam os dados estudantis sob perspectivas diferentes, mas complementares.

Diferentes visualizações, tais como gráficos de barras e pizza, são utilizados pelas ferramentas de acompanhamento citadas. Apenas em uma das instituições existe uma ferramenta desenvolvida que utiliza grafos para visualização de disciplinas e suas interligações, conforme relato de um dos participantes: *"Alguns são bem convencionais, não têm nada de muito mágico ... [para visualização de disciplinas] uso um grafo bem simples"*.

Todas as instituições reportaram problemas relacionados à integração de dados (sistemas isolados, formatos incompatíveis, etc.), sendo que em duas delas, estes problemas eram mais graves, a ponto de demandar muito trabalho de integração a cada período: *"Temos poucos dados e eles estão esparsos. Não se tem uma gerência única de evasão"* (R1, R13).

Durante as entrevistas pôde-se constatar que uma das instituições possui um acompanhamento muito detalhado a respeito de alunos, cursos e disciplinas, disponibilizando para seus gestores *dashboards* elaborados a partir de ferramentas de mercado: *"temos muitos, muitos indicadores ... hoje nós temos um dashboard super completo"*. De forma geral as métricas geralmente acompanhadas são: disciplinas cursadas (número de faltas, aprovação, reprovação, cancelamento); estado no curso (ingresso, reingresso, cancelamento e trancamento); percentual de aprovação (disciplinas cursadas e aprovadas) e o resultados de pesquisas de avaliação ao final de cada semestre (R4, R5, R6).

A movimentação de grupos e de alunos individualmente (transferências internas) não é acompanhada ou é acompanhada de forma limitada, pois, embora considerada importante, não está disponível de forma facilitada (R2, R4). Conforme citado por um participante: *"A visão histórica dos alunos e seu percurso é um desejo"* e não é contemplada apropriadamente nos sistemas hoje utilizados (R1). Além disso, apesar das transferências

[1] Linguagem própria para análises estatísticas [33].

internas serem acompanhadas, esta métrica não é atualmente percebida como ponto de atenção para uma possível condição de evasão, embora todos participantes concordem com a possibilidade deste ser um indicador relevante.

Questionados se o tempo de permanência do aluno no curso poderia ser um indicador de evasão, constatou-se que esta não é uma métrica avaliada em todas instituições, mas se entende como indiretamente relacionada à evasão quando existem prazos para a formação dos alunos (R8). Conforme um participante: *"para nós não é necessariamente um indicador negativo porque o perfil dos nossos alunos é de não fazer uma grade cheia"* e ponderou que esta métrica deve ser contextualizada para cada instituição. A reincidência de disciplinas foi citada espontaneamente como relevante para o acompanhamento geral, mas conforme mencionado, precisa ser contextualizada com o tipo de instituição (ensino público x privado) (R7).

Na maioria dos casos, a ordem com que as disciplinas são cursadas é definida pelo próprio estudante, de acordo com a grade curricular e seus requisitos. Somente uma das instituições possui um trabalho mais ativo no aconselhamento de matrículas. Em uma conversa mais ampla foi considerado que a ordem com que as disciplinas são cursadas (escolha inicial por disciplinas com maior índice de aprovação ou que exijam menor envolvimento, em detrimento das disciplinas do seu semestre) pode ser um indicador de um aluno propenso a uma futura evasão: *"... influenciaria no caso de um aluno que já está em dúvida quanto ao curso"* (R9).

Quando questionados, nem todos os participantes disseram acompanhar diretamente um indicador de evasão, sendo que um deles relatou que: *"nós fomos evoluindo no que seria um indicador de evasão ... hoje acompanhamos a frequência, cancelamentos..."*. Eles também disseram não ter um acompanhamento preditivo para identificar uma possível evasão de forma sistematizada, a ponto de antecipar alguma ação de retenção (R10). Foi citada a possibilidade de identificar alunos que voltaram à instituição após um período de afastamento, embora esta análise não seja feita por todos os participantes.

Com relação ao protótipo que foi apresentado aos participantes, os pareceres foram positivos, principalmente em relação a visão histórica, o que pode ser observado nos seguintes comentários: *"Certamente a gente precisa destes dados e desta visão"*; *"Eu que tenho obsessão por ver as coisas ante ao tempo, ... fico muito satisfeito com essa proposição"*; *"pode ser entendido claramente [a respeito da movimentação de estudantes]"*, *"Ficar catando os dados em diversas fontes é difícil e visualizações assim desse tipo, o cara bate o olho e já vê o que está acontecendo"*.

A facilidade de utilização da interface de visualização foi mencionada, e algumas melhorias foram apontadas, tais como: a suavização do efeito de destaque de nós e ligações; uma melhor descrição de cada nó de disciplina (e não simplesmente uma listagem de alunos) ao usar a sobreposição do mouse; e a possibilidade de outros arranjos de co-

res permitindo o destaque de cada etapa do semestre (ingresso, disciplinas, estados das disciplinas e egresso) (R12).

5.3 Resultados e Requisitos

Os resultados das entrevistas mostram que vários indicadores têm sido acompanhados, mas há interesse numa proposta que permita ter uma visão do percurso de estudantes sob uma perspectiva histórica e integrada, ampliando as possibilidades de exploração analítica. Assim, o resultado da análise das entrevistas corroborou tanto com requisitos identificados na RSL, como com as limitações apontadas.

Nenhum dos participantes tinha conhecimento, a respeito de ferramentas disponíveis em suas instituições, que fornecessem uma visualização de dados como à proposta no protótipo. Portanto, através dos resultados apresentados anteriormente, foi possível definir uma lista de requisitos e oportunidades que podem auxiliar na análise do percurso acadêmico dos estudantes. Estes requisitos estão descritos a seguir e sintetizados na Tabela 5.1.

Tabela 5.1: Requisitos e oportunidades encontradas.

	Requisitos e Oportunidades	Identificado na RSL	Identificado nas Entrevistas	Implementado no Protótipo Preliminar
R1	Visão histórica de percurso	✓	✓	✓ Parcial
R2	Movimentações internas (transferências entre cursos)	✓	✓	✓ Parcial
R3	Acompanhamento geral e de grupos de estudantes	✓	✓	✓ Parcial
R4	Acompanhamento individual de estudantes	✓	✓	✓
R5	Acompanhamento de métricas do curso (formados, evadidos, trancamentos, etc.)	✓	✓	✓ Parcial
R6	Acompanhamento de métricas das disciplinas cursadas (número de faltas, aprovação, reprovação, cancelamento, etc.)	✓	✓	✓ Parcial
R7	Acompanhamento de reincidência de alunos em uma disciplina	-	✓	-
R8	Acompanhamento do tempo de permanência do estudante na instituição	✓	✓	-
R9	Acompanhamento da ordem de realização das disciplinas	-	✓	-
R10	Acompanhamento de alunos com tendência a abandonarem seus cursos	✓	✓	-
R11	Filtros (por curso, forma de ingresso, forma de egresso, movimentações, por aluno)	✓	✓	✓ Parcial
R12	Interface fácil de usar e interagir	✓	✓	✓ Parcial
R13	Ferramental e modelo, abertos para uso nas instituições	✓	✓	-

- **R1 - Visão histórica de percurso:** Presente nos trabalhos da RSL, foi mencionada nas entrevistas como sendo um desejo, assim como a possibilidade de visualização

de múltiplos cursos ao mesmo tempo, que também foi uma necessidade constatada por O'Handley et al. [50]. Neste sentido, a visualização baseada em diagrama de *Sankey* se mostrou adequada para os dados deste estudo.

- **R2 - Movimentações internas (transferências):** Presente nos trabalhos da RSL, embora não seja acompanhada enfaticamente nas instituições dos participantes da pesquisa, pode complementar o estudo de evasão por meio da visualização de fluxos de cursos distintos concomitantemente.
- **R3 - Acompanhamento geral e de grupos de estudantes:** Presente nos trabalhos da RSL e entrevistas, este acompanhamento geral e de grupos é comumente feito nas instituições de ensino, segundo as entrevistas, seja a partir de uma visão geral do semestre ou de agrupamentos como turma, curso, forma de ingresso, etc.
- **R4 - Acompanhamento individual de estudantes:** Identificado como uma necessidade na RSL e também a partir das entrevistas, é importante para permitir um olhar individualizado a respeito de indicadores do alunos, permitindo que ações sejam tomadas de acordo com a criticidade dos mesmos.
- **R5 - Acompanhamento de métricas do curso:** Presente nos trabalhos da RSL e mencionado nas entrevistas, deve permitir o acompanhamento do progresso dos alunos a partir do seu ingresso na instituição como um novo aluno (vestibular, transferência externa, reingresso), passando pelo curso escolhido (curso 1, curso 2, curso 3), até seu resultado ao final do semestre (formado, transferência interna, evadido, matriculado, trancado).
- **R6 - Acompanhamento de métricas das disciplinas cursadas:** Assim como cursos, a possibilidade de acompanhamento de disciplinas realizadas e seu desempenho (aprovado, reprovado, taxa de aprovação, cancelamento, número de faltas) foi citada em trabalhos da RSL e nas entrevistas.
- **R7 - Acompanhamento de reincidência de alunos em uma disciplina:** Mencionado nas entrevistas, pode ser usado para identificar tanto alunos com alguma dificuldade, quanto disciplinas que precisam ser readequadas em um curso.
- **R8 - Acompanhamento do tempo de permanência do estudante na instituição:** Citado nas entrevistas como um parâmetro de acompanhamento importante e presente na literatura analisada [31, 28, 40]. Segundo um dos especialistas ouvidos, esta métrica deve ser delimitada ao contexto de cada instituição, pois é comum alunos não concluírem seus cursos dentro do tempo esperado (referindo-se a sua instituição).
- **R9 - Acompanhamento da ordem de realização das disciplinas:** Mencionado nas entrevistas e visto como uma oportunidade, pode apontar situações nas quais

disciplinas são postergadas por algum motivo específico, levando a um atraso na grade curricular.

- **R10 - Acompanhamento de alunos com tendência a abandonarem seus cursos:** Presente em alguns trabalhos da RSL e mencionado nas entrevistas, diz respeito à implementação de forma sistemática de um modelo preditivo de evasão com base nas movimentações de alunos pelos semestres.
- **R11 - Filtros:** Presente em alguns trabalhos da RSL, a disponibilidade de filtro por período (seleção de certo intervalo de tempo) foi citada como uma necessidade por um entrevistado, assim como outras possibilidades de filtro relacionadas aos dados apresentados (curso, forma de ingresso, forma de egresso, movimentações, por aluno, etc.).
- **R12 - Interface fácil de usar e interagir:** Algumas interações e detalhes da interface foram identificadas na RSL ou foram mencionadas nas entrevistas como pontos de melhoria do protótipo, como por exemplo: efeito de destaque de arestas (as linhas permanecem em cor, enquanto os elementos que não estão diretamente ligados esmaecem); outros arranjos de cores; possibilidade de arrastar nós para uma melhor visualização; destaque por cor de nós críticos (reprovados, evadidos, baixa média de aprovação) e detalhamento de nós a partir de *tooltips*.
- **R13 - Ferramental e modelo abertos para uso nas instituições:** Embora não seja possível resolver o problema de integração de dados internos em diferentes instituições, é importante disponibilizar um ferramental e modelo de dados que permita o uso da visualização com dados educacionais independente da instituição.

Os requisitos identificados representam algumas das necessidades dos especialistas entrevistados e ajudaram a compor o modelo de análise visual descrito no próximo capítulo.

6. DESCRIÇÃO E IMPLEMENTAÇÃO DO MODELO PROPOSTO

A partir dos requisitos levantados e apresentados na Seção 5.3, foi projetado um modelo visual composto por um conjunto de técnicas de visualização interativas com objetivo de permitir a exploração e análise de dados acadêmicos ao longo do tempo. Os detalhes deste modelo são descritos na Seção6.1. A implementação, que consiste em uma evolução do protótipo apresentado no Capítulo 4, é descrita na Seção 6.2, enquanto as visualizações são apresentadas na Seção 6.3. O modelo de predição, criado a partir do emprego de técnicas estatísticas, é discutido na Seção 6.4.

6.1 Modelo de Visualização

Modelos podem ser definidos como uma representação parcial de uma ideia, objeto, evento ou processo, que é produzida com objetivos específicos [26]. Neste trabalho, é proposto um modelo para representação de um fluxo de estudantes ao longo de semestres acadêmicos.

O projeto do modelo foi amparado pelos requisitos apresentados na Seção 5.3. A partir da análise do protótipo desenvolvido e apresentado aos especialistas de domínio, o modelo proposto tenta complementar as lacunas identificadas, assim como trazer novas possibilidades em relação ao acompanhamento de estudantes.

A Figura 6.1 representa o modelo visual com seus diferentes componentes, que deve permitir a visualização e a interação com as duas principais entidades acadêmicas: alunos e disciplinas.

O componente de filtros (componente 1 na Figura 6.1) corresponde aos vários tipos de seleções que devem ser possíveis no conjunto de dados. Através deste componente, é atendido o requisito R11, além de viabilizar uma série de análises a partir da seleção de um aluno ou de um agrupamento de alunos filtrados por um determinado quesito, contemplando os requisitos R2 a R10. Ao aplicar qualquer filtro, deve ser gerado um subconjunto de dados para uso no modelo, e consequentemente, todos os demais componentes visuais do modelo deverão ser sensibilizados (componentes 2, 3, 4a e 5), juntamente com o modelo preditivo (componente 6).

O resumo do conjunto de dados (componente 2) deve apresentar uma visão rápida sobre detalhes gerais do coorte selecionado e dos seus principais indicadores na forma de totalizadores. Desta forma, é possível atender aos requisitos R3, R4 e R5 (quando selecionado um estudante, disciplina ou curso em específico).

Os componentes 3 e 4a definem a apresentação e exploração de detalhes do conjunto de alunos e disciplinas do agrupamento, enquanto o componente 4b define uma

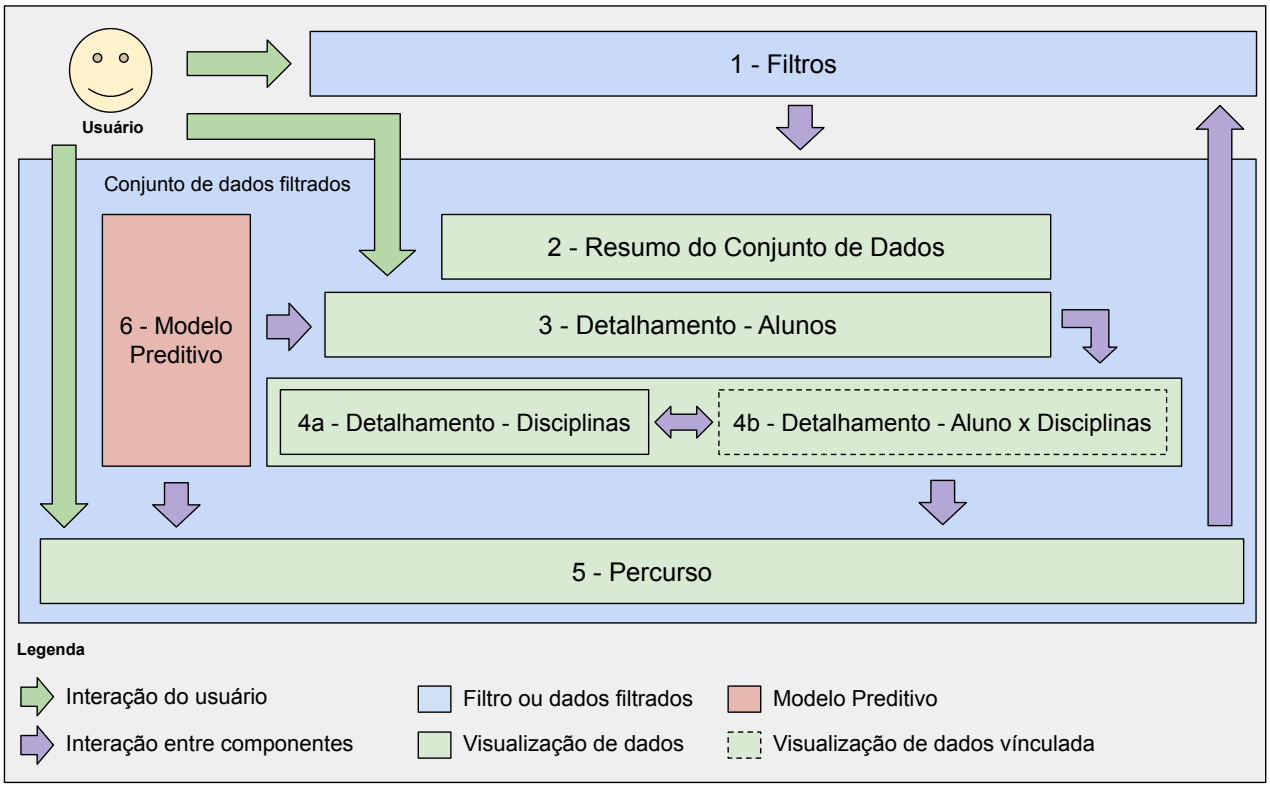


Figura 6.1: Modelo de visualização proposto.

visão particular das disciplinas de um único aluno, selecionado previamente através do componente 3, e atendendo aos requisitos R3, R4 e R6. Ainda no componente 3, devem ser disponibilizadas informações dos estudantes, permitindo atender aos requisitos R3, R4, R7, R8 e R10. Tanto o componente 3, quanto o componente 4b, devem permitir o acompanhamento de alunos reincidentes, atendendo ao requisito R7.

A visualização de percurso propriamente dita, é representada no componente 5 e atende aos requisitos de R1 a R10. O modelo também prevê algumas interações de usuário representadas pelas setas verdes na Figura 6.1, tais como o uso de filtros (componentes 1) e a navegação em profundidade ou vinculada (*drilldown*) por disciplinas de um aluno, a partir da seleção do mesmo no componente 3. As setas roxas nesta figura, demonstram interações entre os componentes do modelo, como, por exemplo, quando um aluno é selecionado no componente 3, o componente 4a deve ser substituído por uma visualização mais detalhada representada pelo componente 4b (que por sua vez, deve acionar um filtro local na visualização de percurso - componente 5). Segundo Chen et al. [16], esta navegação vinculada acontece quando uma seleção, em uma visualização anterior, leva a uma segunda na qual a estética de “visibilidade” é usada para codificar o grau de interesse, como no caso, apenas as linhas das disciplinas relevantes devem ser mostradas.

Na visualização de percurso, por meio da interação por clique em um nó ou ligação, deve ser possível sensibilizar o filtro de forma a criar um novo agrupamento de

alunos. Esta funcionalidade, que conecta a visualização de percurso ao filtro (e, consequentemente, aos demais componentes visuais do modelo), é chamada de filtro cruzado (*crossfilter*) [73].

O modelo preditivo (componente 6) é sensibilizado pelo dados filtrados e está conectado aos componentes de detalhamento de alunos (componente 3) e à visualização do percurso (componente 5).

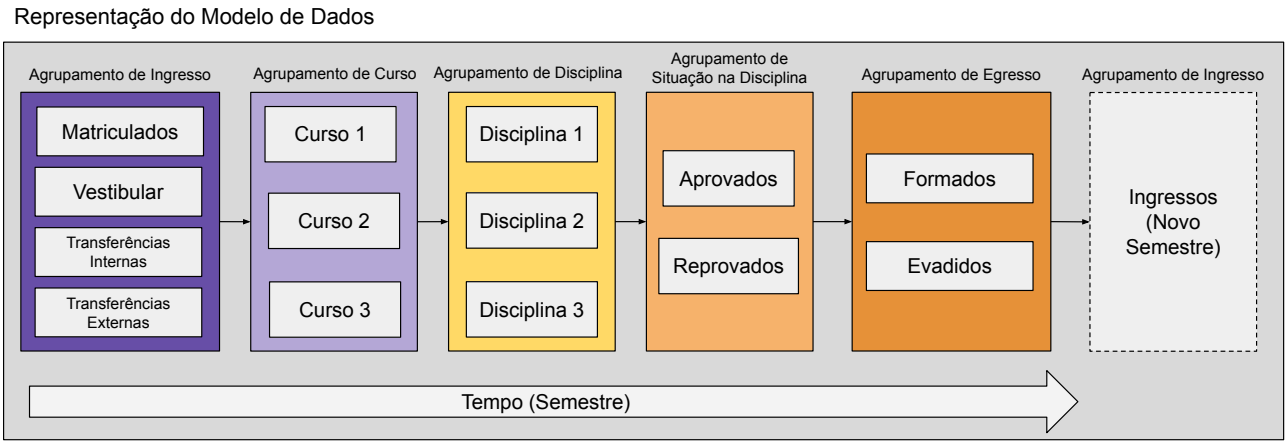


Figura 6.2: Modelo de dados para a visualização de percurso acadêmico.

Para a visualização de percurso, um novo modelo de dados (em relação ao arranjo utilizado no protótipo) foi proposto para atender os requisitos de visão simultânea de múltiplos cursos (R1). Este modelo, apresentado na Figura 6.2, é composto por cinco agrupamentos (Ingresso, Curso, Disciplina, Situação na Disciplina e Egresso) representando as etapas do percurso de um estudante ao longo de um semestre. Foi definido um esquema de cores padrão para o modelo, no qual o roxo representa o início do semestre, o amarelo representa o andamento (disciplinas) e o laranja representa o final, a partir das formas de egresso e demais movimentações.

Cada um dos agrupamentos possui nós descrevendo os estados possíveis e ligações representando as movimentações de alunos em um determinado momento do semestre. Ao término de cada semestre, os agrupamentos finais (Situação na Disciplina e Egresso) se conectam aos próximos estados dos agrupamentos iniciais do semestre seguinte, caso o aluno permaneça na instituição (Matriculados ou Transferências Internas). Este comportamento é demonstrado no destaque da implementação pelas bordas serrilhadas na Figura 6.2.

Cabe mencionar que este modelo de dados foi limitado ao conjunto de dados usado neste estudo, mas não deve ser considerado restrito conceitualmente, pois facilmente pode incorporar dimensões e métricas adicionais (de acordo com a implementação e dados disponíveis). Ainda sobre o modelo de dados, devido às limitações do conjunto de dados utilizado neste estudo, a métrica de número de faltas, relacionada às disciplinas (R6), não pôde ser incluída. Também relacionadas às disciplinas, algumas formas

de saída (cancelou, desistiu, reprovado por média, reprovado por faltas) presentes no agrupamento “Situação na Disciplina” apresentado no protótipo, foram arranjadas em um único estado de “Reprovados” na implementação, com o intuito de manter o modelo de dados independente e permitir atender ao requisito R13.

Nas próximas seções são apresentados detalhes de como este modelo foi implementado e os requisitos atendidos.

6.2 Dicionários de dados e tecnologias utilizadas

Tanto a implementação do modelo como do protótipo, compartilham o mesmo conjunto de dados e a mesma arquitetura de acesso já descritos, respectivamente, nas Seções 4.1 e 4.2). Portanto, disponibilizam dados em uma interface baseada na web, utilizando uma API para realizar consultas a uma base de dados relacional SQL. Estas consultas são previamente organizadas para representarem os filtros do modelo e para realizarem os agrupamentos responsáveis pelos cálculos de indicadores (totalizadores, médias, diferenças, etc.).

Um pequeno grupo de dimensões de dados, descritos previamente no dicionário de dados na Tabela 4.1, possibilitou a derivação de um conjunto completo de variáveis de alunos e de disciplinas. De forma simplificada, estas variáveis nada mais são do que métricas e dimensões compostas por agrupamentos e seleções do próprio conjunto de dados, como, por exemplo, a quantidade de semestres cursados ou total de disciplinas aprovadas, e que foram identificadas por meio da literatura e pelas entrevistas com especialistas. Os detalhamento destas variáveis pode ser visto nos dicionários de dados das Tabelas 6.2 (Alunos) e 6.1 (Disciplinas).

Dimensão ou métrica	Tipo	Descrição
semestredadisciplina	TEXTO	Semestre da disciplina (Ex: 2019/1)
codigodadisciplina	TEXTO	Identificador da disciplina (Ex: Disciplina 1234)
cursos	TEXTO	Nome dos cursos em que a disciplina está vinculada (ex: Curso 1,Curso 2)
alunos_total	INTEIRO	Total de alunos na disciplina (ex: 2019/1)
aprovados_total	INTEIRO	Total de alunos aprovados na disciplina (ex: 2019/1)
reprovados_total	INTEIRO	Total de alunos reprovados na disciplina (ex: 2019/1)
taxaaprovacao	DECIMAL	Relação entre os alunos aprovados e que cursaram a disciplina (ex: 0.99)
taxaaprovacao_geral	DECIMAL	Relação entre os alunos aprovados e que cursaram a disciplina independente de filtros (ex: 0.99)

Tabela 6.1: Dicionário de dados das dimensões e métricas de disciplinas usadas na implementação do modelo, com tipo e descrição.

A interface visual foi desenvolvida utilizando a biblioteca de código aberto chamada *Streamlit*[1], que é baseada na linguagem de programação Python[2] e particularmente

[1] https://streamlit.io/
[2] https://www.python.org/

Identificação	Tipo	Descrição	Modelo
codigodapessoa	INTEIRO	Identificador do aluno (ex: 1234)	-
codigodaturma	TEXTO	Código da turma no formato [id_curso]-[semestre] (ex: 1-2016/2)	-
curso	TEXTO	Nome do curso (ex: Curso 1)	-
curso_final	TEXTO	Nome do curso após transferência interna (ex: Curso 1)	-
transferencia_interna	BOOLEANO	Realizou transferência interna de curso (ex: 1)	INDEP.
curso_final_bacharelado	BOOLEANO	Curso final é do tipo bacharelado (ex: 1)	INDEP.
semestredeingresso	TEXTO	Primeiro semestre na instituição (ex: 2019/1)	-
ultimosemestrematriculado	TEXTO	Último semestre na instituição (ex: 2019/1)	-
semestredeingresso_transferencia	TEXTO	Primeiro semestre após uma transferência interna (ex: 2019/1)	-
ultimosemestrematriculado_transferencia	TEXTO	Último semestre após uma transferência interna (ex: 2019/1)	-
formadeingresso_regular	TEXTO	Forma de ingresso (ex: Vestibular)	-
formadeegresso_regular	TEXTO	Forma de egresso (ex: Formado)	-
ingresso_regular	BOOLEANO	Realizou ingresso por meio de Vestibular (ex: 1)	INDEP.
disciplinas	INTEIRO	Número total de disciplinas cursadas (ex: 23)	INDEP.
disciplinas_reincidencias	INTEIRO	Número total de disciplinas cursadas (ex: 3)	INDEP.
disciplinas_aprovado	INTEIRO	Número total de disciplinas cursadas aprovadas (ex: 20)	INDEP.
disciplinas_reprovado	INTEIRO	Número total de disciplinas cursadas reprovadas (ex: 3)	INDEP.
media_disciplinas_semestre	INTEIRO	Média de disciplinas por semestre (ex: 2)	-
total_semestres	INTEIRO	Número total de semestres cursados (ex: 6)	INDEP.
taxa_conclusao	DECIMAL	Relação entre o número de disciplinas aprovadas e o total necessário para a conclusão do curso (ex: 0.61)	INDEP.
taxa_conclusao_norm	INTEIRO	Taxa de conclusão normalizada para inteiro (ex: 61)	INDEP.
taxa_aprovacao	DECIMAL	Relação entre o número de disciplinas aprovadas e o número de disciplinas realizadas (ex: 0.99)	INDEP.
taxa_aprovacao_norm	INTEIRO	Taxa de aprovação normalizada para valores inteiros (ex: 99)	INDEP.
formacao_atrasada	BOOLEANO	Formação excedeu o limite de semestres padrão do curso (Ex: 1)	INDEP.
formadeegresso_final	TEXTO	Forma de egresso após transferência interna (ex: Formado)	-
evadiu	BOOLEANO	Aluno evadiu da instituição (ex: 1)	DEPEN.
evasao_prob	DECIMAL	Probabilidade de evasão do aluno de acordo com o modelo preditivo (ex: 0.99)	RESUL.

Tabela 6.2: Dicionário de dados com as dimensões e métricas de alunos usadas na implementação do modelo, com tipo e descrição. A coluna Modelo indica sua utilização como variável no modelo preditivo de acordo com o seu tipo: Independente (INDEP.), Dependente (DEPEN) ou Resultado (RESUL.)

usada para a criação de aplicações de ciência de dados e aprendizado de máquina [65]. Os gráficos interativos foram criados através da biblioteca *Plotly*[3], e a biblioteca *matplotlib*[4] foi usada para os gráficos mais específicos, como o de matriz de confusão e mapa de calor, apresentados na área de modelo preditivo (Subseção 6.4.1). Para o diagrama de *Sankey* foi utiliza a mesma biblioteca *Echarts*, mencionada anteriormente na Seção 4.2. A manipulação interna de dados foi feita utilizando a biblioteca *Pandas*[5], que fornece estruturas de dados de alto desempenho e de fácil manipulação. Para o modelo esta-

tístico foram utilizadas as bibliotecas *statsmodels*[6] e *scikit-learn* [7], apropriadas para este trabalho e amplamente utilizadas na área de ciência de dados e afins.

6.3 Visualizações e Funcionalidades

Nesta seção são descritas a visualização e as funcionalidades desenvolvidas. Segundo Ferreira et al. [24], independentemente da visualização a ser criada, sua principal função é oferecer maior facilidade ao usuário na execução de determinada tarefa e, assim, para atender a todos os seus requisitos.

A partir dos requisitos, a implementação foi feita com o objetivo de apresentar o modelo visual de uma forma que permitisse uma rápida e fácil navegação (atendendo ao requisito R12) por uma série de dados detalhados. Neste sentido, os trabalhos de Vaclavek et al. [72] e Ferreira et al. [24], propondo o uso, com sucesso, de *Learning Analytics Dashboards* para o acompanhamento de alunos e seus diferentes indicadores, serviram de inspiração para o desenvolvimento.

A Figura 6.3, ilustra a tela principal da implementação, com a área de Acompanhamento exibida, desenvolvida a partir do modelo proposto e responsável por incorporar os principais requisitos identificados. Outras áreas ajudam a compor a implementação: Detalhamento Gráfico, Modelo Preditivo e Conjunto de Dados. Acessíveis pelo menu de abas (Fig.6.3B), estas áreas são descritas em detalhes nas seções a seguir.

6.3.1 Área de Acompanhamento

A área de Acompanhamento contém a implementação do modelo de visualização, sendo a principal área para realização das análises. É composta por um menu lateral (*sidebar*), aberto pelo ícone na lateral superior esquerda, que dá acesso aos filtros do conjunto de dados (Figura 6.3A). Estes filtros, conforme descrito na Seção 6.1, permitem que diferentes coortes sejam criadas em tempo real, por meio das inúmeras seleções possíveis. A Tabela 6.3 apresenta a relação completa das opções de filtro e suas descrições.

O resumo estatístico geral (Figura 6.3C) apresenta uma visão concisa do conjunto de dados atual (filtrados ou não), possibilitando um rápido olhar sobre o todo de forma quantitativa. Os indicadores disponíveis são: período (intervalo de tempo), cursos, semestres, disciplinas, alunos, evadidos, formados, transferências internas, matriculados,

[3] https://plotly.com/
[4] https://matplotlib.org/
[5] https://pandas.pydata.org/
[6] https://www.statsmodels.org/
[7] https://scikit-learn.org/

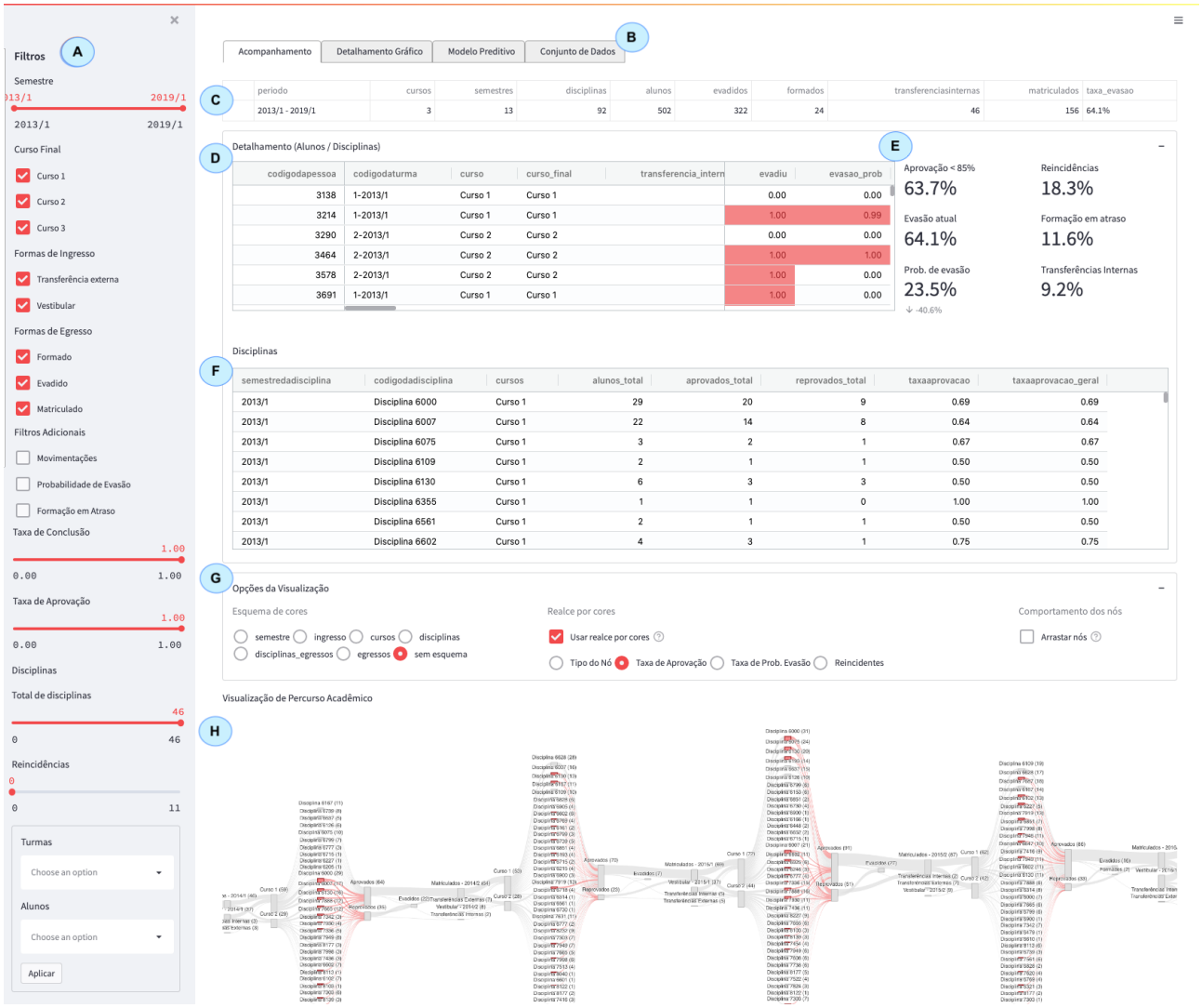


Figura 6.3: Tela principal da implementação com seus componentes relacionados ao modelo visual, destacados por marcadores em azuis com letras de A a H.

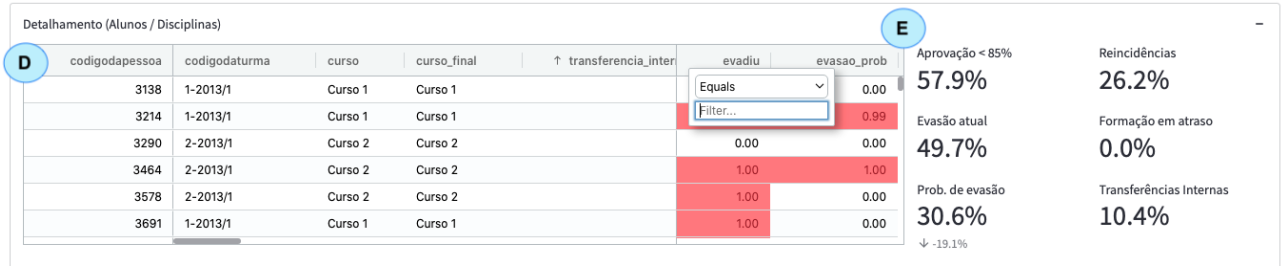


Figura 6.4: Destaque da tela de Acompanhamento de alunos.

taxa de evasão. Este resumo pode ser usado em conjunto com os filtros no sentido de apresentar resultados de grupos específicos de alunos, por curso, período, entre outros.

Diferentemente da visão de resumo geral, o detalhamento de alunos (Figura 6.3D) e o detalhamento de disciplinas (Figura 6.3F), possibilitam uma visão ampliada em formato tabular de indicadores de todos os alunos e disciplinas do agrupamento. A utilização de tabelas e resumos estatísticos é ideal nos casos nos quais é necessária a visualiza-

Nome	Tipo	Descrição
Semestre	Controle Deslizante	Intervalo de semestres do conjunto de dados.
Curso Final	Caixa de Seleção	Último curso (após transferência interna ou não) frequentado.
Forma de Ingresso	Caixa de Seleção	Forma ingresso do aluno (Transferência externa, Vestibular).
Forma de Egresso	Caixa de Seleção	Forma atual de egresso do aluno (Formado, Evadido, Matriculado).
Filtros Adicionais - Movimentações	Caixa de Seleção	Alunos que fizeram uma transferência interna de cursos.
Filtros Adicionais - Probabilidade de Evasão	Caixa de Seleção	Usa uma taxa de probabilidade evasão acima de 50%, atribuída pelo modelo de predição, para a seleção de alunos.
Filtros Adicionais - Formação em Atraso	Caixa de Seleção	Filtra alunos com um número de semestres maior que o previsto pelo seu curso final.
Taxa de Conclusão	Controle Deslizante	Taxa de conclusão (número de disciplinas aprovadas em relação ao total necessário para a formação) igual ou menor ao valor selecionado.
Taxa de Aprovação	Controle Deslizante	Taxa de aprovação (relação entre o número de disciplinas realizadas e aprovadas) igual ou menor ao valor selecionado.
Disciplinas - Total de disciplinas	Controle Deslizante	Valor igual ou maior ao número de disciplinas realizadas no total.
Disciplinas - Reincidências	Controle Deslizante	Valor igual ou maior ao número de disciplinas reincidentes.
Turmas	Caixa Combo	Uma ou mais turmas usando o seu identificador único.
Alunos	Caixa Combo	Um ou mais alunos usando o seu identificador único.

Tabela 6.3: Descrição das possíveis opções de filtro.

ção de dados exatos, segundo Chen at al. [16]. Para facilitar a navegação pelos dados, o componente de visualização das tabelas permite rearranjo e ordenação por coluna e filtros pontuais por valor. Um detalhe ampliado deste componente pode ser visto na Figura 6.4D.

Complementando a visão detalhada, o resumo estatístico de alunos (Figura 6.4E), apresenta alguns indicadores de desempenho como: taxa de evasão (atual), taxa de alunos com probabilidade de evasão (e diferença em relação a evasão atual), taxa de alunos com uma média de aprovação inferior a 85%, taxa de alunos com alguma reincidência em disciplinas, taxa de alunos que ultrapassaram o número de semestres estabelecido para o curso, taxa de alunos com alguma transferência interna de curso.

Ainda sobre a tabela de detalhamento de alunos, além dos diversos indicadores apresentados (Tabela 6.2), esta visão permite que um aluno seja selecionado e mais infor-

Figura 6.5: Destaque da tela de Acompanhamento de disciplinas de um aluno.

mações possam ser exploradas a seu respeito. A Figura 6.5 demonstra um possível caso de uso, no qual é analisado o percurso de um estudante que realizou uma movimentação de curso (destacado em roxo no diagrama G), através de uma transferência interna (destaque em laranja) e teve uma reprovação em uma disciplina (destaque em vermelho), posteriormente recuperada (reincidência). Esta possibilidade de acompanhamento de alunos reincidentes atende ao requisito R7. Também pela imagem, é possível observar os detalhamentos das disciplinas cursadas (Figura 6.5F1), visão simplificada de disciplinas aprovadas e reprovadas ao longo do tempo (Figura 6.5F2) e sua movimentação acadêmica no período (Figura 6.5G), como já citada, por meio do diagrama de *Sankey*. As dimensões

e métricas disponíveis neste detalhamento por disciplinas estão descritas no dicionário de dados da Tabela 6.1.

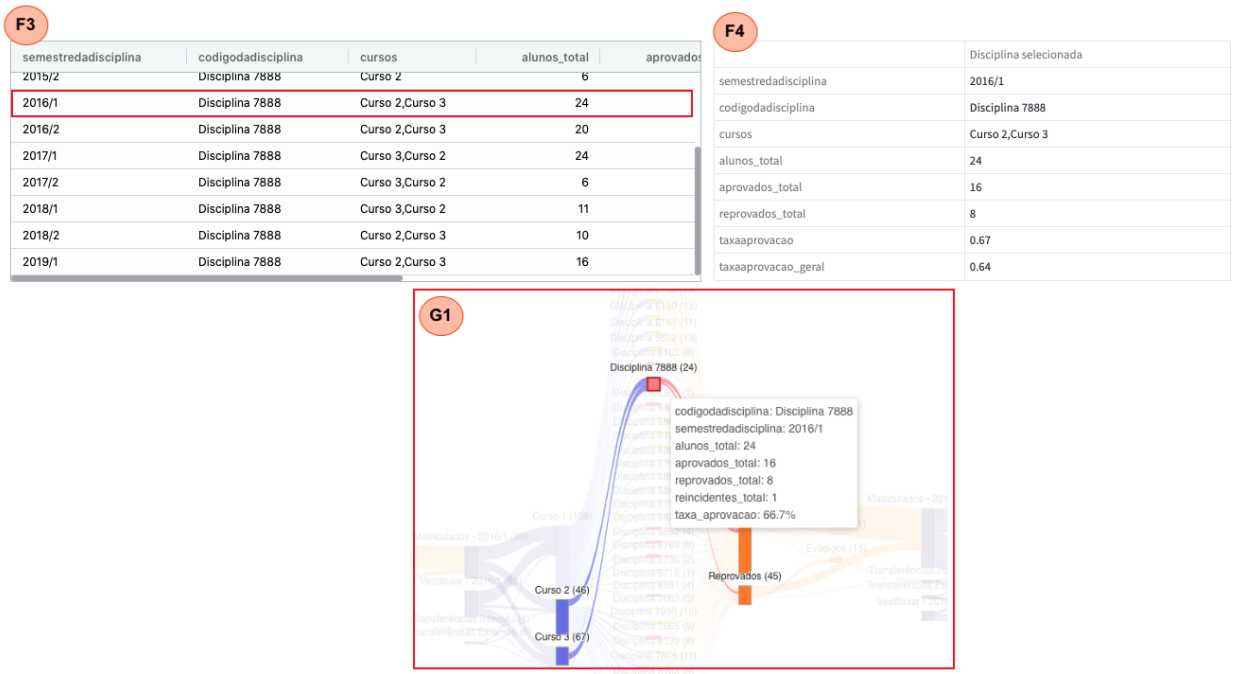


Figura 6.6: Destaque de uma disciplina ao longo de um período (F1) e detalhes da seleção feita pelo diagrama de percurso (F2). Detalhes de uma disciplina em um *tooltip* do diagrama (G1) - a cor vermelha na disciplina representa um percentual de menos de 85% de aprovações.

Por fim, a visualização de percurso acadêmico (Figura 6.4H) e suas opções (Figura 6.4G), é o principal componente do modelo, e que possibilita a visão do “todo”, permitindo uma rápida exploração interativa pelo conjunto de dados e se conectando aos demais componentes de detalhamentos. Foi desenvolvida a partir do protótipo apresentado no Capítulo 4 e acrescida das novas funcionalidades alinhadas aos requisitos levantados. Por meio da implementação (Figura 6.7) do novo modelo de dados, apresentado na Figura 6.2, são possíveis diferentes análises muti-curso de forma simultânea, ideal para o acompanhamento das movimentações de alunos.

Além disso, foram adicionadas novas interações, como o *crossfilter* (descrito na Seção 6.1) e o detalhamento de uma disciplina ao clicar em um nó correspondente (Figura 6.6F1 e 6.6F2). A quantidade de alunos únicos pode ser vista em um “tooltip”, ao usar o movimento de “pairar sobre” / *hover* do mouse sobre um nó ou ligação de forma geral, enquanto mais detalhes são apresentados caso o nó represente uma disciplina, conforme ilustra a Figura 6.6G1. Os rótulos dos nós apresentam uma descrição e o número de alunos únicos do grupo (esta grandeza também pode ser percebida e comparada pela altura de cada nó e a espessura das ligações entre eles). O diagrama ainda permite a navegação panorâmica e *zoom*, que podem ser acionados pelo movimento de arrastar e soltar, e pelo botão de rolagem, respectivamente.

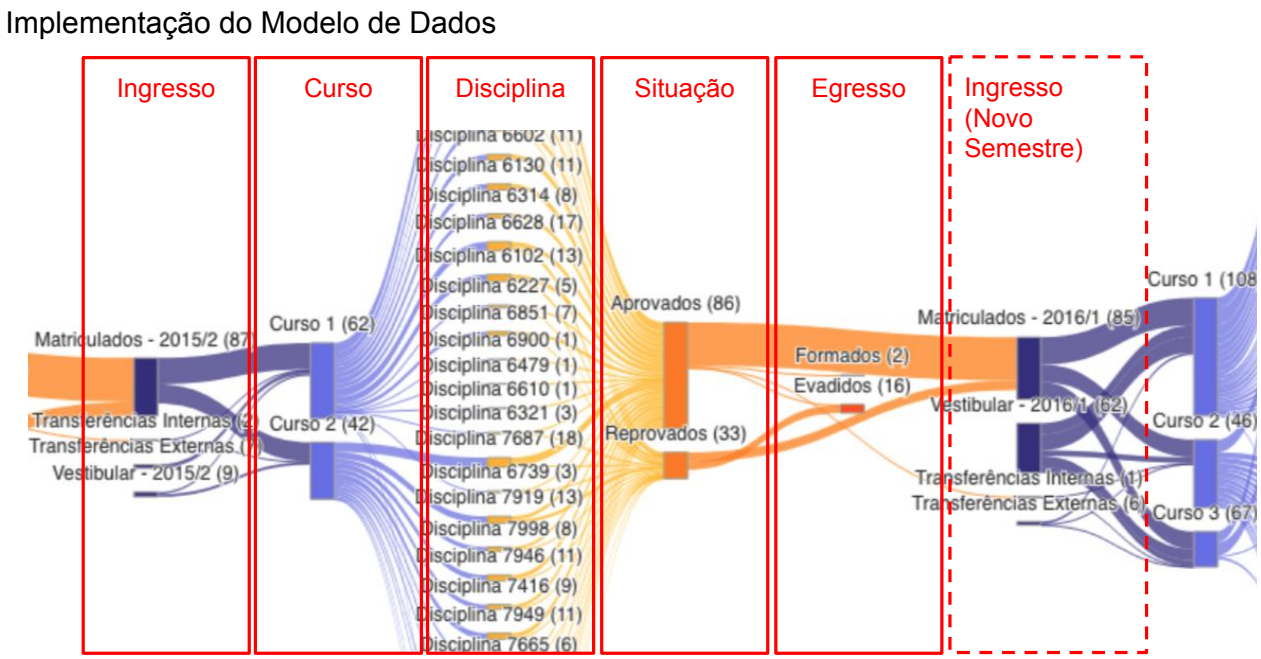


Figura 6.7: Descrição da relação entre o modelo de dados (Figura 6.2) e a implementação.

Para aprimorar a capacidade de análise através do diagrama de percurso, algumas opções de ajustes desta visualização estão disponíveis e podem ser vistas na interface apresentada na Figura 6.4G. Estas opções são:

- **Esquema de cores** - Permite a alteração das cores usadas nos diferentes agrupamentos de dados apresentados na visualização, para que sejam destacadas áreas de interesse do usuário, tais como semestre, ingresso, cursos, disciplinas, situação na disciplina, egresso ou ainda a possibilidade de remover o esquema de cores.

- **Realce por cores** - Tem a finalidade de permitir uma rápida visualização de estudantes de um grupo em risco, por meio do realce por cores nos agrupamentos e nós do diagrama, a partir de critérios selecionáveis como: tipo de nós, taxa de aprovação na disciplina, probabilidade de evasão e reincidências na disciplina.

- **Comportamento dos nós** - Permite a livre movimentação dos nós do diagrama de *Sankey*, que por padrão são fixados para que a navegação panorâmica seja possível pela interação por mouse.

Estas facilidades de interação entre os componentes, diferentes esquemas e realces por cores, detalhamentos em geral, etc., estão relacionadas ao requisito R12 (Interface fácil de usar e interagir).

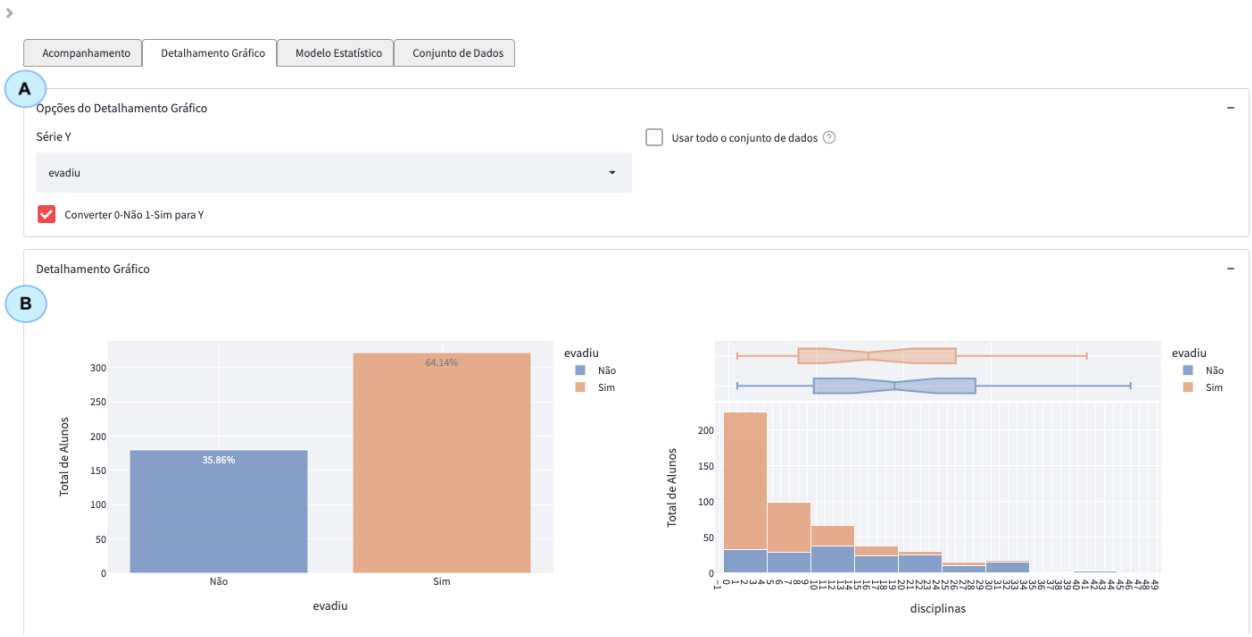


Figura 6.8: Tela parcial do Detalhamento Gráfico. O destaque (A) são as opções disponíveis. O destaque (B) apresenta dois (dos 12) possíveis gráficos comparativos - Total de alunos x Evasão (esquerda) e Total de Alunos x Disciplinas x Evasão (direita)

6.3.2 Área de Detalhamento Gráfico

Ao explorar um conjunto de dados, pode-se obter informações sobre as quais não se tinha conhecimento inicialmente. Este paradigma de exploração é chamado de “análise exploratória de dados” - EDA (*Exploratory Data Analysis*). No processo de EDA, partindo dos dados, procuram-se pistas que levem a conjecturas, testando e retrocedendo, até que algum “insight” valioso seja descoberto. Esta tarefa é especialmente complexa quando muitas variáveis (dimensões e métricas) estão envolvidas, e é neste momento que a visualização de dados pode ajudar [3].

Visando possibilitar uma análise exploratória visual dos dados, a área de detalhamento gráfico exibe um conjunto de gráficos que permite fazer uma série de comparações de indicadores associados ao desempenho de estudantes. Além disso, permite analisar livremente relacionamentos entre atributos, que antes não haviam sido mapeados.

A Figura 6.8 apresenta um recorte da interface da área de detalhamento gráfico. A organização da área é feita em três blocos: opções (Figura 6.8A), detalhamento (Figura 6.8B) e gráfico customizado (Figura 6.9). O processo de EDA ocorre através do acompanhamento das variáveis pré-selecionadas e sensíveis ao conjunto de dados atual (com ou sem filtro). Os gráficos de barras empilhadas e histogramas (com detalhamento de valores marginais, mediana, mínimo e máximo) possibilitam que as variáveis sejam comparadas com, por exemplo, o número de evasões ou outra variável selecionável.

Os gráficos pré-selecionados de detalhamento disponíveis nesta área são:

- Evadiu x Total de Alunos (Gráfico de Barras);
- Disciplinas x Total de Alunos x Evadiu (Sim, Não) (Histograma Empilhado);
- Transferência Interna x Total de Alunos x Evadiu (Sim, Não) (Gráfico de Barras Empilhado);
- Ingresso Regular x Total de Alunos x Evadiu (Sim, Não) (Gráfico de Barras Empilhado);
- Curso Final x Total de Alunos x Evadiu (Sim, Não) (Gráfico de Barras Empilhado);
- Média de Disciplinas por Semestre x Total de Alunos x Evadiu (Sim, Não) (Histograma Empilhado);
- Total de Semestres x Total de Alunos x Evadiu (Sim, Não) (Histograma Empilhado);
- Taxa de Aprovação x Total de Alunos x Evadiu (Sim, Não) (Histograma Empilhado);
- Taxa de Conclusão x Total de Alunos x Evadiu (Sim, Não) (Histograma Empilhado);
- Formação atrasada x Total de Alunos x Evadiu (Sim, Não) (Gráfico de Barras Empilhado);
- Curso Final Bacharelado x Total de Alunos x Evadiu (Sim, Não) (Gráfico de Barras Empilhado);
- Disciplinas Reincidências x Total de Alunos x Evadiu (Sim, Não) (Gráfico de Barras Empilhado).

No componente de opções (Figura 6.8A) é possível alterar a variável de comparação (eixo y) padrão dos gráficos do detalhamento, optar pela conversão ou não de rótulos binários e se os gráficos serão gerados usando o conjunto de dados atual (filtrado) ou todo o conjunto de dados.

Os gráficos destacados na Figura 6.8B representam dois possíveis gráficos comparativos. O primeiro, no lado esquerdo, é um gráfico de barras que mostra o Total de alunos x Evasão, apresentando o total de alunos evadidos (laranja), que é de 64,14% neste exemplo. O segundo gráfico, no lado direito, consiste em um histograma que compara a distribuição do número total de disciplinas realizadas em relação a situação de evasão (Total de Alunos x Disciplinas x Evasão). Pela exploração visual é possível identificar que grande parte dos alunos evadidos para este conjunto de dados não cursou mais de quatro disciplinas.

Como última funcionalidade nesta área, pela interface, é possível criar um gráfico customizado (Figura 6.9) para permitir outras explorações não identificadas previamente.

Figura 6.9: Detalhamento Gráfico - Tela apresentando a possibilidade de criação de gráficos customizados por meio da seleção dos eixos x e y, e tipo de visualização.

6.3.3 Área de Conjunto de dados

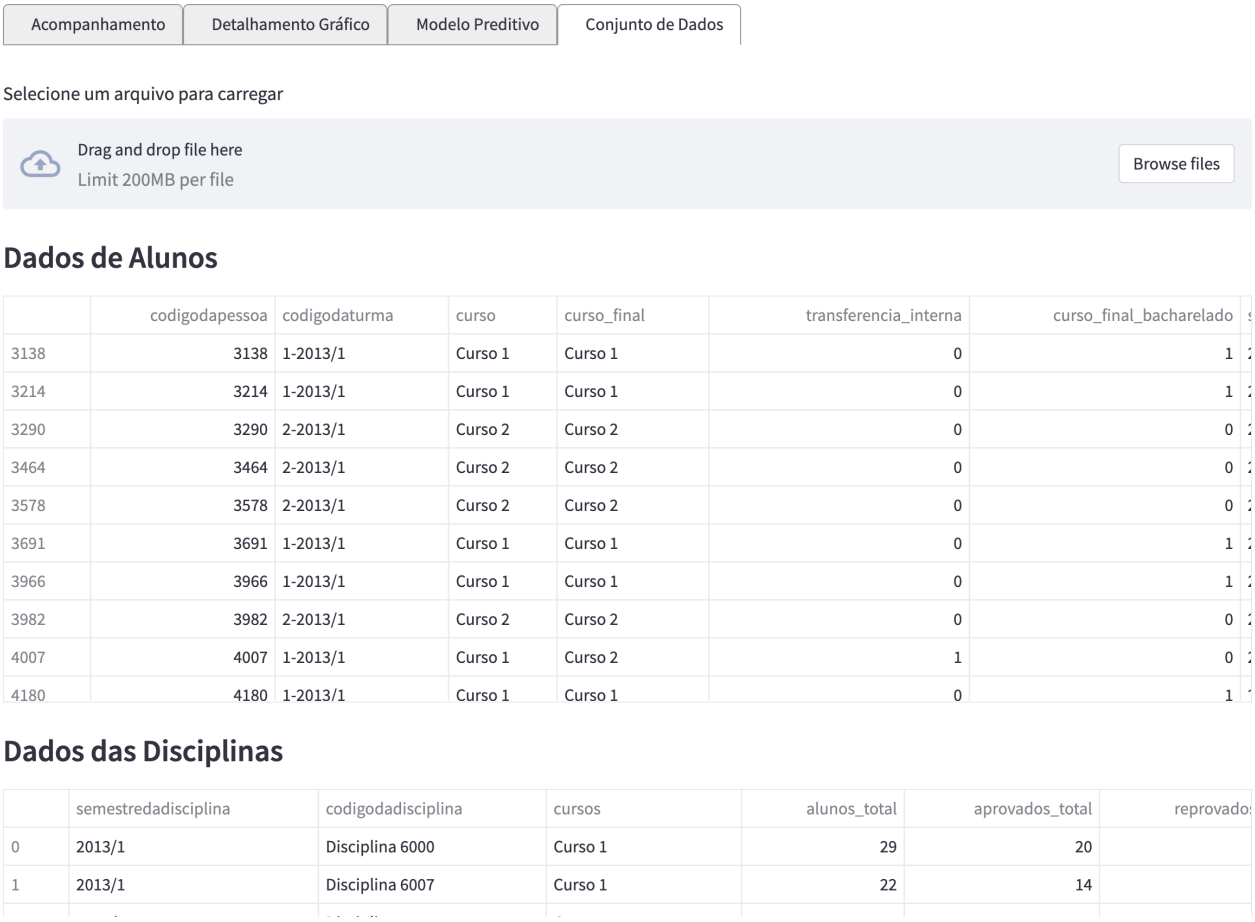


Figura 6.10: Visão da interface de área de Conjunto de dados

A área chamada de conjunto de dados foi implementada visando atender ao requisito R13, que consiste na possibilidade de uso do modelo e sua implementação com dados das diferentes instituições. Portanto, nesta área é feito o gerenciamento dos dados brutos a serem usados, isto é, a partir desta interface, dados próprios das instituições, seguindo o arranjo de dimensões descrito no dicionário de dados da Tabela 4.1, poderiam ser enviados para uso. A Figura 6.10 mostra parte da interface com as áreas de carga de dados e a apresentação do conjunto de dados brutos existente (alunos e disciplinas). Esta separação (alunos e disciplinas) ocorre a partir do processamento do arquivo de dados enviado previamente.

6.4 Modelo preditivo

Conforme identificado, a possibilidade de utilização de um modelo preditivo em conjunto da visualização foi citada como muito relevante para apoiar o acompanhamento de estudantes. O objetivo da predição é inferir um atributo de destino (variável dependente) por meio da combinação de diferentes atributos deste mesmo conjunto de dados (variáveis independentes) [58]. Desta forma, é possível calcular a probabilidade desta variável dependente possuir um determinado estado, como por exemplo, se um aluno irá ou não evadir de uma instituição. Métodos de predição são ditos de classificação quando a variável prevista é um valor categórico binário (ex:Sim ou Não, 1 ou 0), ou categórico multinomial (ex: Formado, Matriculado e Evadido) [58, 41].

Para a construção do modelo preditivo foram seguidos os passos da metologia de mineração de dados CRISP-EDM (*CRoss-Industry Standard Process for Educational Data Mining*), adaptada e proposta por Ramos et al. [55] e própria para a área educacional. A Figura 6.11 ilustra este modelo.

As etapas adaptadas e seguidas neste trabalho foram:

1. **Entendimento dos domínios educacionais da evasão** - A dimensão de evasão (retenção) foi escolhida e caracterizada como um problema a ser abordado, por ser um tema recorrente nos trabalhos analisados [32, 37, 2, 19] e nas entrevistas com especialistas, devido a sua criticidade para as instituições e indicadores educacionais.

2. **Entendimento do conjunto de dados disponível** - Por meio dos dados disponibilizados pela instituição de ensino (que foram previamente verificados em busca de inconsistências, como valores nulos e atípicos), uma série de variáveis descritas no dicionário de dados da Tabela 6.2, delimitadas pela coluna modelo e identificadas pelos valores Independente (INDEP.), Dependente (DEPEN.), Resultado (RESUL.), foram criadas para que fossem possíveis as análises e a construção do modelo.

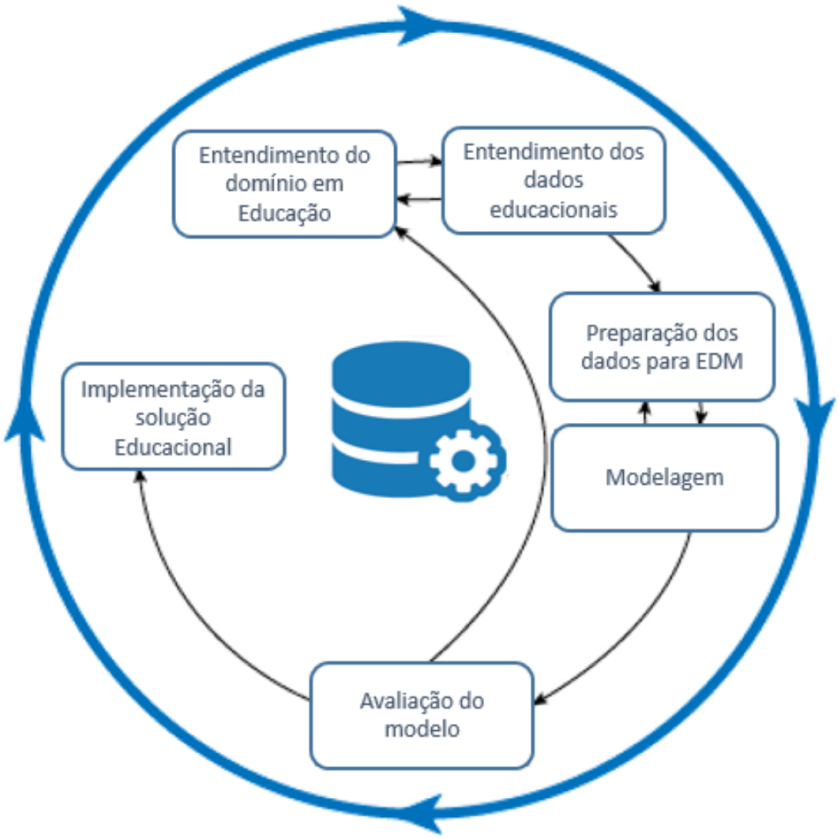


Figura 6.11: Modelo do CRISP-EDM. Fonte: Ramos et al. [55]

3. **Preparação dos dados para EDM** - Nesta etapa foram investigados, através de um processo visual de análise exploratória de dados (EDA) e utilizando a área de detalhamento gráfico (Subseção 6.3.2), padrões relacionados ao objetivo do modelo para que fossem selecionadas variáveis representativas e que ajudassem na sua construção.

4. **Definição do modelo preditivo** - A partir do uso da técnica de aprendizagem supervisionada chamada Regressão Logística, foram criados modelos preditivos da evasão do aluno, com base nas variáveis obtidas a partir do processo de EDA. Em EDM, a técnica de Regressão Logística tem sido usada para prever o desempenho de alunos por meio de um conjunto de variáveis independentes numéricas ou categóricas (multivariada) [54, 2]. Esta técnica, além de apresentar bons resultados, possui as vantagens de não requer um relação linear entre a variável dependente e as variáveis independentes, e ser menos afetada, quando suposições estatísticas básicas, como a normalidade dos dados[8], não são satisfeitas [63].

5. **Avaliação do modelo preditivo** - Com base nas métricas de qualidade de modelos preditivos da literatura [52] (descritas na Subseção 6.4.1), e direcionados pelo princípio da simplicidade de *Occam* [21], cada variação do modelo gerado foi analisada até a escolha de uma representação adequada para a predição da evasão.

[8]Que segue uma distribuição normal.

6. **Implementação da solução Educacional** - O modelo preditivo criado e avaliado (pela etapa anterior), ajudou a compor a implementação do modelo visual descrita neste trabalho. Posteriormente, o conjunto foi submetido a uma avaliação por meio de entrevistas à especialistas da área de educação.

As próximas Subseções apresentam, o processo seguido para a criação do modelo preditivo e disponibilizado em uma área própria da implementação, assim como as etapas técnicas de construção e avaliação do modelo em questão.

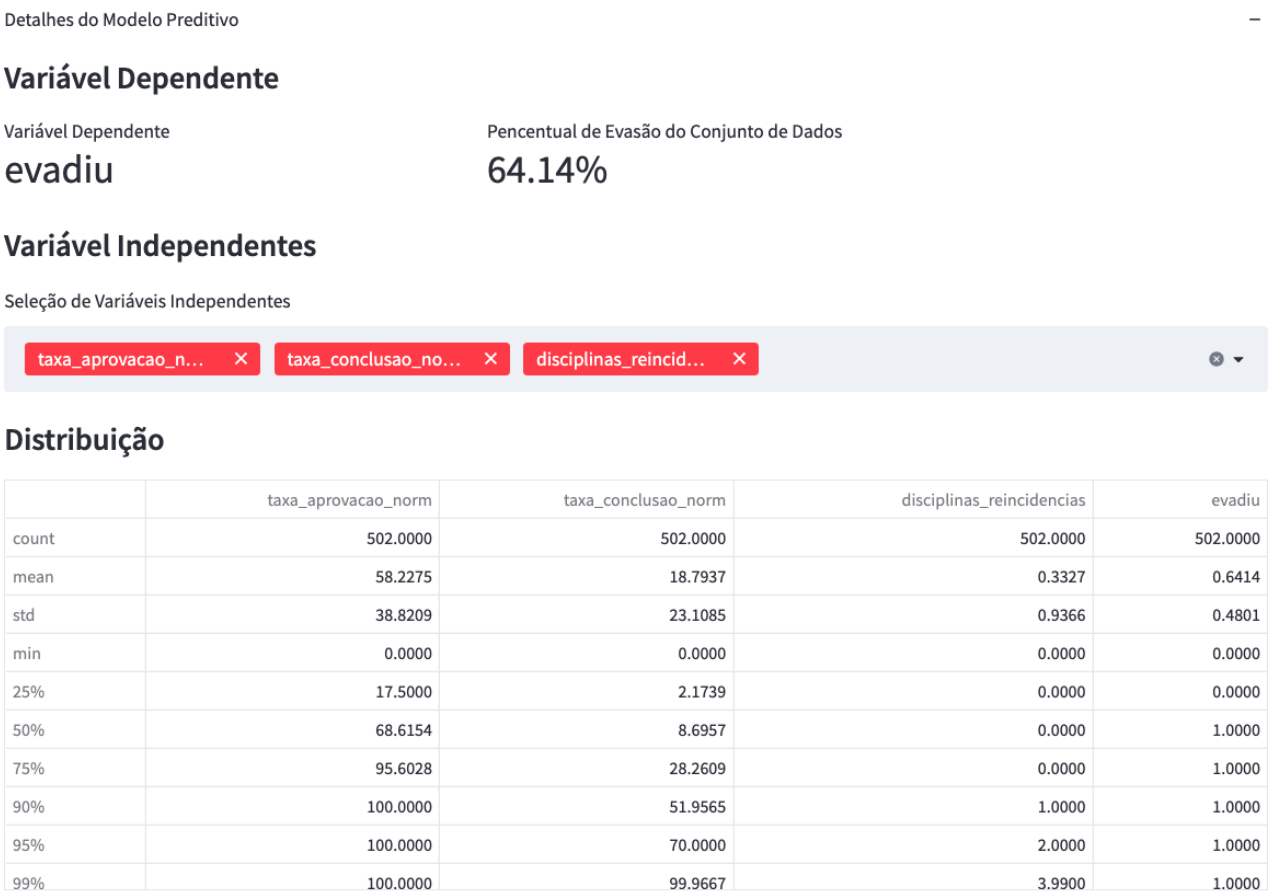
6.4.1 Área de Modelo Preditivo

De forma a tornar o processo de definição do modelo preditivo mais claro e flexível a outros cenários e conjuntos de dados, uma área contendo o conjunto de passos usado na construção deste modelo foi desenvolvida e disponibilizada na implementação. Nesta área, acessível pela aba “Modelo Preditivo” (Figura 6.3B), além de ser possível acompanhar visualmente as etapas de criação do modelo, alguns parâmetros podem ser modificados de acordo com necessidades identificadas e aplicados em tempo real, alterando o resultado final da predição. A seguir são descritas as etapas de construção do modelo de predição e sua relação com os elementos disponibilizados pela interface.

- **Variáveis do modelo** - Nesta etapa são selecionadas as variáveis independentes que serão utilizadas para a predição da variável dependente (evadiu). Pela interface é possível a escolha de outras variáveis independentes e sua aplicação ao modelo para posterior avaliação. A Figura 6.12 traz um visão da interface associada a esta etapa, incluindo uma visualização tabular da descrição dos dados, contendo indicadores como: número de registros, valor médio de cada variável, valor mínimo e máximo, desvio padrão e percentis (25%, 50%, 75%, 90%, 95%, 99%). Esta descrição dos dados, é útil para identificar eventuais problemas como a presença de valores nulos ou atípicos (*outliers*).

- **Correlação de Variáveis** - Utilizando diferentes testes (*Pearson, Kendall, Spearman*) selecionáveis pela interface, é possível identificar a intensidade e direção (de acordo com o sinal do valor) do coeficiente da correlação entre as variáveis independentes e a variável dependente do modelo. Segundo Chok [17], os coeficientes de correlação de *Pearson*, *Spearman* e *Kendall*, são as medidas mais comumente usadas em testes de associação monotônicas[9], com os dois últimos geralmente sugeridos para dados não distribuídos normalmente. Na Figura 6.13 é possível observar o

[9]Uma relação monotônica se dá quando uma variável aumenta ou diminui continuamente à medida que outra variável aumenta [62]



	taxa_aprovacao_norm	taxa_conclusao_norm	disciplinas_reincidencias	evadiu
count	502.0000	502.0000	502.0000	502.0000
mean	58.2275	18.7937	0.3327	0.6414
std	38.8209	23.1085	0.9366	0.4801
min	0.0000	0.0000	0.0000	0.0000
25%	17.5000	2.1739	0.0000	0.0000
50%	68.6154	8.6957	0.0000	1.0000
75%	95.6028	28.2609	0.0000	1.0000
90%	100.0000	51.9565	1.0000	1.0000
95%	100.0000	70.0000	2.0000	1.0000
99%	100.0000	99.9667	3.9900	1.0000

Figura 6.12: Interface da etapa de variáveis do modelo contendo: a indicação da variável dependente, percentual desta ao longo do conjunto de dados, seleção de variáveis independentes e visualização tabular da descrição dos dados.

teste de correlação e suas visualizações disponíveis: gráfico de mapa de calor e de forma tabular.

Já o teste de multicolinearidade por *Variance Inflation Factor* (VIF) é usado para mensurar o grau de inter-relacionamento entre variáveis independentes. Desta forma espera-se antecipar as variáveis com forte multicolinearidade que poderiam causar instabilidade no modelo. O valor de corte adotado foi de 10, pois, de forma geral, os valores de VIF superiores a este são considerados como de forte inter-relacionamento [64], porém, este limiar pode ser ajustado através da interface (Figura 6.13). De acordo com o resultado do teste VIF e do limiar escolhido, variáveis independentes são listadas em uma visão tabular, que sugere sua remoção.

- **Divisão do conjunto de dados entre treino e teste** - Como uma prática comum no desenvolvimento de modelos [69], convêm a separação do conjunto de dados em parte teste e parte treino, para que o modelo de aprendizado não seja influenciado por todos os valores de uma só vez. Os valores padrões definidos foram de 70% do conjunto de dados para o treino do modelo e os demais 30% foram utilizados para os testes de validação. Estes valores podem ser configurados pela interface e os

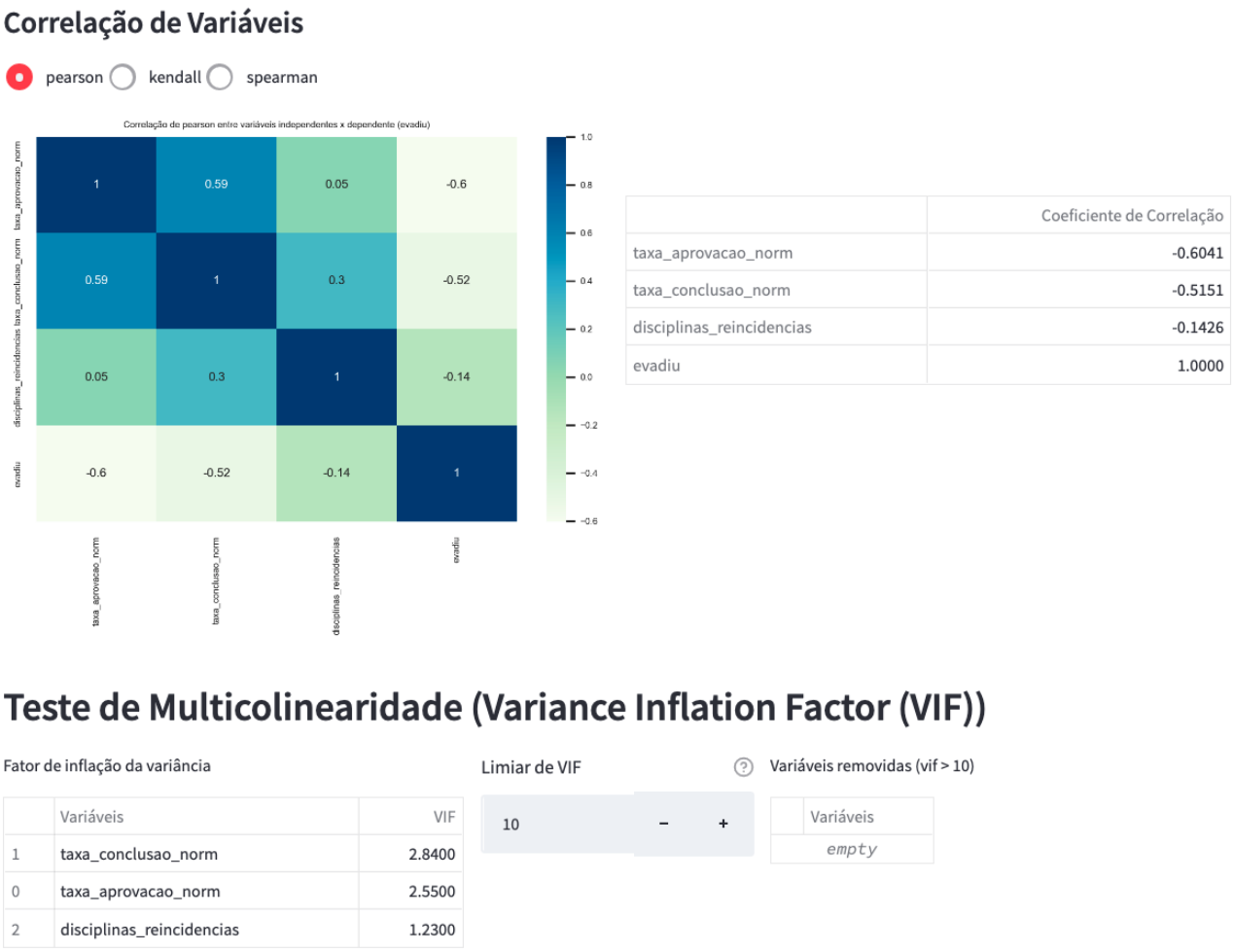


Figura 6.13: Interface da etapa de correlação de variáveis: possibilita a escolha de diferentes testes de coeficientes de correlação e apresenta o teste de multicolinearidade por VIF.

conjuntos de dados resultantes são descritos em formato tabular, conforme ilustra a Figura 6.14.

- **Modelo de Regressão Logística** - Nesta etapa o modelo é executado e uma visão tabular, gerada pela biblioteca *statsmodel*, é apresentada. Esta visão inclui indicadores como os coeficientes de regressão (coef), que dão uma ideia da variação observada no modelo para cada unidade incrementada da variável, além da direção desta variação (positiva ou negativa) [23]. Outro indicador observado é o valor-P, usado para acompanhamento da significância das variáveis independentes no modelo [25]. A Figura 6.15 apresenta a visão tabular de resumo do modelo descrita e uma representação da fórmula[10] de Regressão Logística aplicada.

- **Métricas de Avaliação** - Como a maioria dos modelos são falíveis, diferentes escolhas de parâmetros fornecerão diferentes taxas de falsos positivos (FP - *false positive*) e falsos negativos (FN - *false negative*), assim como de classificações verdadei-

[10] Fórmula utiliza a notação padrão R - (wilkinson-rogers) [51].

Divisão do Conjunto de Dados entre Treino e Teste

Percentual da divisão para Testes

0.30

0.00 1.00

Conjunto de Dados - Treino

	taxa_aprovacao_norm	taxa_conclusao_norm
count	351.0000	351.0000
mean	59.4016	19.3960
std	38.2609	23.7195
min	0.0000	0.0000
25%	26.1364	2.1739
50%	70.0000	8.6957
75%	95.8333	30.4348
max	100.0000	100.0000

Conjunto de Dados - Teste

	taxa_aprovacao_norm	taxa_conclusao_norm
count	151.0000	151.0000
mean	55.4983	17.3938
std	40.0885	21.6331
min	0.0000	0.0000
25%	0.0000	0.0000
50%	66.6667	8.6957
75%	93.7500	25.5435
max	100.0000	100.0000

Figura 6.14: Interface da etapa de divisão do conjunto de dados entre treino e teste. Através da interface a proporção padrão (70% / 30%) pode ser alterada. De forma complementar, é possível uma visão tabular da descrição destes conjuntos de dados.

Fórmula do Modelo Preditivo

```
evadiu ~ taxa_aprovacao_norm + taxa_conclusao_norm + disciplinas_reincidencias
```

Sumário do Modelo Logístico

Dep. Variable:	evadiu	**No. Observations:**	502
Model:	GLM	**Df Residuals:**	498
Model Family:	Binomial	**Df Model:**	3
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-190.62
Date:	Fri, 08 Apr 2022	**Deviance:**	381.24
Time:	13:17:22	**Pearson chi2:**	722.
No. Iterations:	7	**Pseudo R-squ. (CS):**	0.4206
Covariance Type:	nonrobust		

Generalized Linear Model Regression Results

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	5.8820	0.659	8.931	0.000	4.591	7.173
taxa_aprovacao_norm	-0.0637	0.008	-8.165	0.000	-0.079	-0.048
taxa_conclusao_norm	-0.0189	0.007	-2.786	0.005	-0.032	-0.006
disciplinas_reincidencias	-0.4508	0.155	-2.899	0.004	-0.756	-0.146

Figura 6.15: Interface da etapa de execução do modelo de Regressão Logística, com uma visão tabular de sumário e destacando os indicadores de coeficiente de regressão e valor-P. Na parte de cima na figura, uma representação da fórmula do modelo.

ras (TP - *true positive* e TN - *true negative*). Este acompanhamento pode ser feito por meio da “Matriz de Confusão”, disponível como um gráfico de mapa de calor para os valores do conjunto de dados de treino e teste [52].

Já a Curva ROC (*Receiver Operating Characteristic*) e sua área abaixo da curva - **ROC-AUC** (*Area Under Curve*) são uma representação visual (gráfico de área) das taxas de verdadeiros positivos (TPR) em relação às taxas de falsos positivos para todos os limites possíveis entre 0 e 1. O melhor modelo produz uma área sob a curva (ROC-AUC) com valor próximo a 1, representando sua eficiência no processo de classificação [52]. Na Figura 6.16 pode ser vista a representação gráfica das matrizes de confusão (treino e teste) e da curva ROC.

Outras métricas relacionadas às taxas de acertos e erros que são comumente utilizadas e estão disponíveis para mensuração da qualidade do modelo, são: Acurácia, Precisão, Revocação e Medida-F1. Segundo Rahman e Devanbu [52]:

- **Acurácia** do modelo é a proporção de previsões corretas (TP e TN), calculada pela Fórmula 6.1.

$$\text{Acurácia} = \frac{TP + TN}{TP + FP + TN + FN} \tag{6.1}$$

- **Precisão** indica a quantidade de resultados positivos corretos dentre o total de positivos previstos para determinada categoria. É calculada pela Fórmula 6.2.

$$\text{Precisão} = \frac{TP}{TP + FP} \tag{6.2}$$

- **Revocação** ou ***Recall***, indica o total de positivos encontrados entre o total real de positivos na amostra para determinada categoria e é calculada pela Fórmula 6.3. O modelo com baixo valor de revocação seria incapaz de encontrar a maioria dos defeitos.

$$\text{Revocação} = \frac{TP}{TP + FN} \tag{6.3}$$

- **Medida-F1** ou ***F1-score*** *ou F1* definida como a média harmônica entre precisão e revocação, calculada pela Fórmula 6.4. Como existe uma relação entre as métricas de precisão e revocação, que pode levar um valor a subir em detrimento do outro (distorcendo a avaliação), a métrica F1 é utilizada para mitigar estes indicadores.

$$\text{Medida-F1} = \frac{2 * \text{Precisão} * \text{Revocação}}{\text{Precisão} + \text{Revocação}} \tag{6.4}$$

O método chamado de validação cruzada estratificada (*Stratified K-Folds*), que procura qualificar a aplicação do modelo por meio de sucessivas rodadas de validação feitas em diferentes segmentos do conjunto de dados de treino e teste, foi usado como base para composição das métricas mencionadas. A variação “estratificada” foi escolhida para manter o uso da razão de separação de dados definida (70% para treino e 30% para testes) [56].

As métricas mencionadas, incluindo ROC-AUC, podem ser vistas na Figura 6.16 em formato tabular e organizadas de forma a comporem uma visão histórica. Para cada alteração dos parâmetros do modelo de predição, como a inclusão de uma nova variável independente, é possível acompanhar este novo conjunto de métricas e compará-las às do modelo anterior.

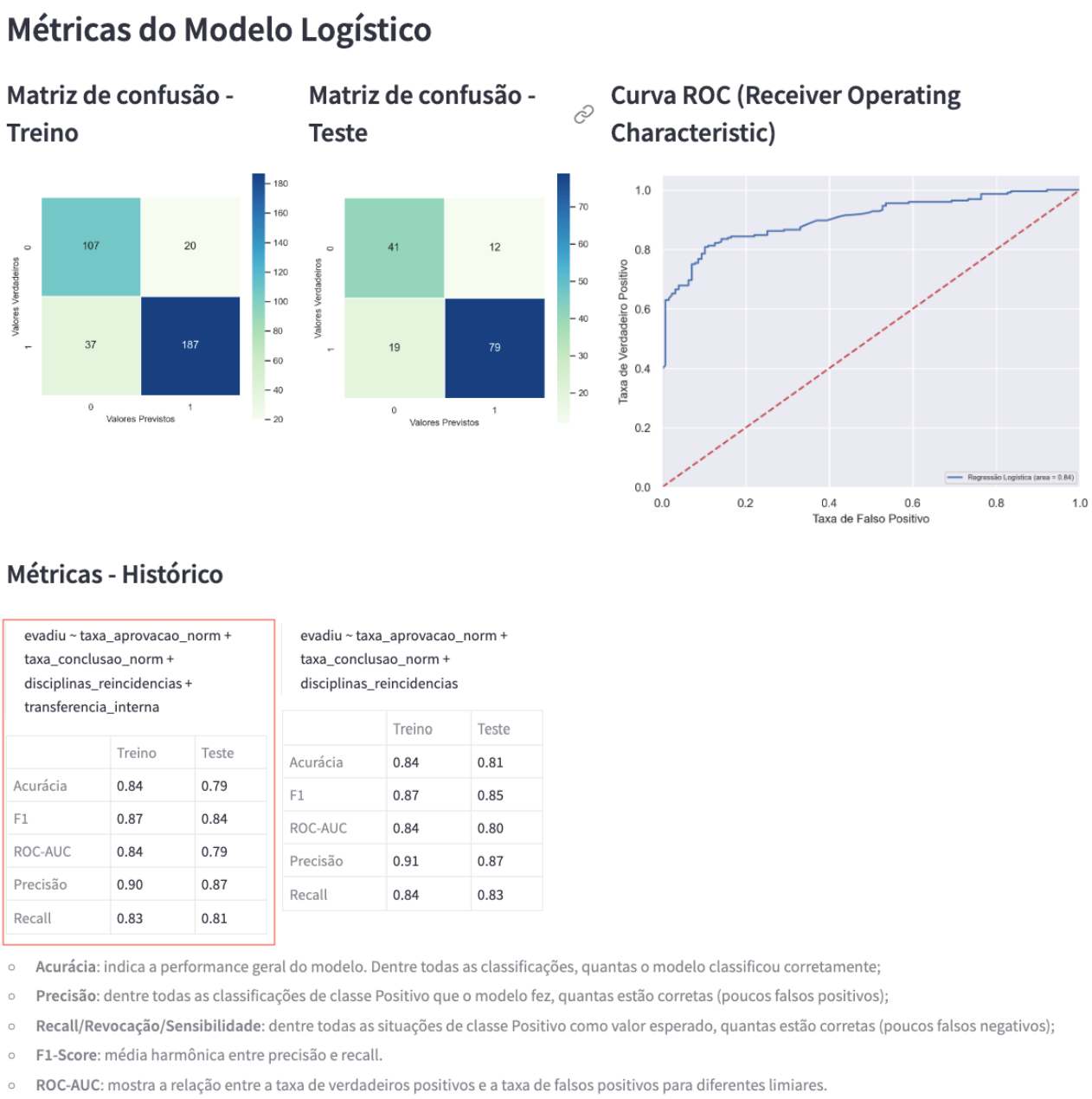


Figura 6.16: Interface da etapa de Métricas de Avaliação, com a representação gráfica das matrizes de confusão para os conjuntos de dados de treino e teste, um gráfico de área representando a curva ROC do modelo e uma visão tabular das métricas de avaliação. O destaque (vermelho) representada uma versão atual do modelo, podendo ser comparada a uma versão anterior (com diferente parametrização).

- **Simulação de modelo** - Na interface de simulação podem ser testados diferentes cenários relacionados aos indicadores educacionais do modelo para um aluno fictício

e obter sua probabilidade de evasão. Esta funcionalidade do tipo *what-if*[11], embora, inicialmente pensada para os tomadores de decisão, pode eventualmente ajudar em ações de orientação diretamente apresentadas aos alunos, como sugerem Gubbala et al. [28].

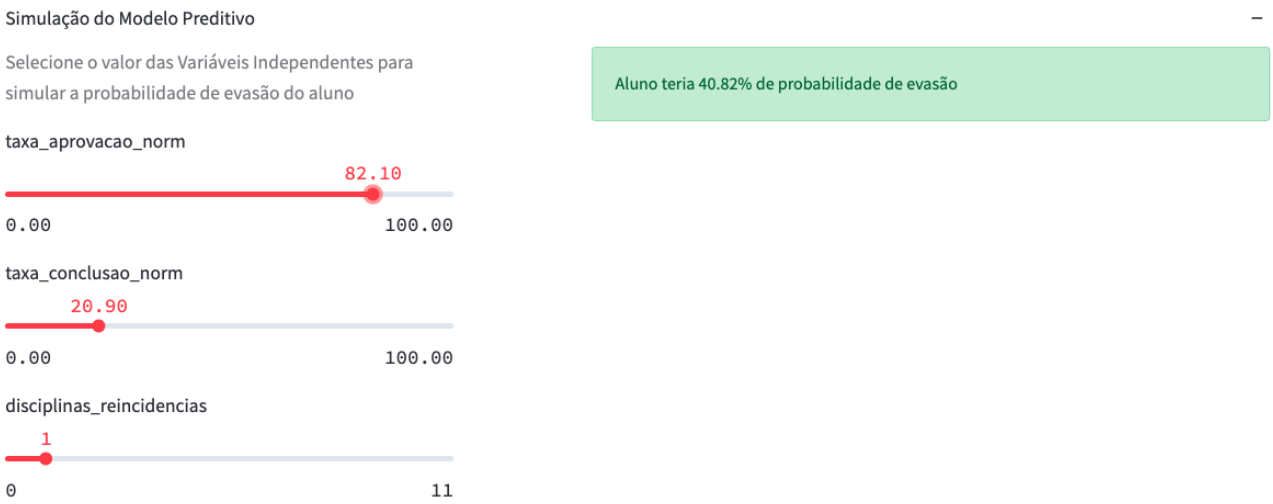


Figura 6.17: Interface de simulação do modelo preditivo, que possibilita testar cenários, através da alteração de indicadores de desempenho usados no modelo, e visualizar a probabilidade de evasão resultante.

6.4.2 Construção e Análise do Modelo Preditivo

Para a definição do modelo final utilizado na implementação, foi realizado um processo chamado de seleção de variáveis. Segundo Guyon et al. [29], existem muitos benefícios potenciais advindos desta etapa, como facilitar a visualização e compreensão de dados, reduzindo o tempo de treinamento de um modelo e, principalmente, a dimensionalidade, de forma melhorar o desempenho de modelos de predição. Dentre as abordagem possíveis para realizar esse processo, a escolhida, devido a sua simplicidade e eficiência, foi a seleção por fator de multicolinearidade (VIF) [1]. Adicionalmente, após a seleção inicial, o modelo preditivo foi rodado algumas vezes para que fossem identificadas e excluídas as variáveis com menor significância (valor-P) [23].

A Figura 6.18 ilustra esse processo de seleção de variáveis, que foi conduzido da seguinte forma: através de EDA, foram selecionadas variáveis independentes que poderiam ser relevantes para a criação do modelo preditivo e que posteriormente foram submetidas a um processo de avaliação. A partir dos critérios de multicolinearidade (VIF), cada variável com valores fora de conformidade para o modelo foi sendo removida até que uma versão final do conjunto fosse obtida. Utilizando este conjunto, uma nova etapa

[11] *E se* - no sentido de simulação de cenários [74].

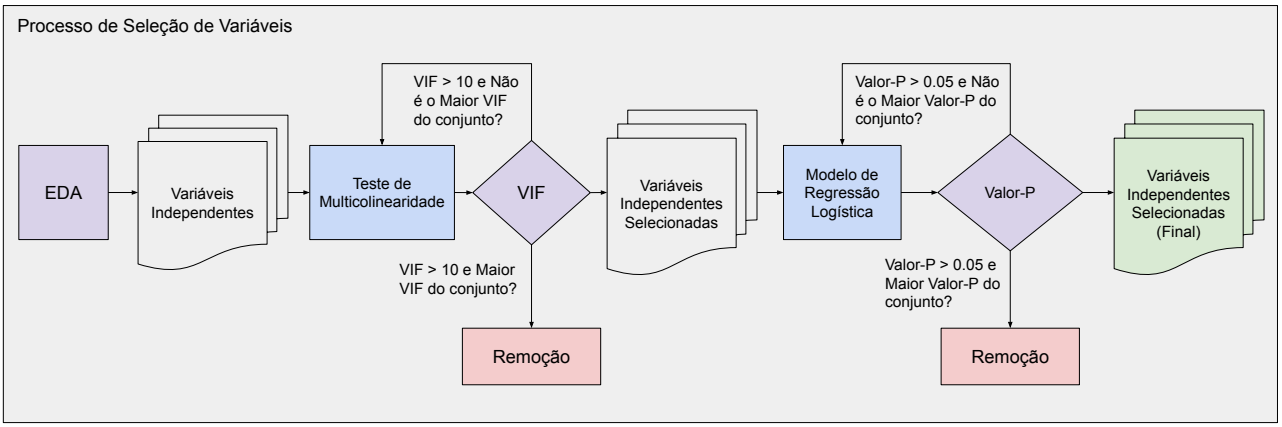


Figura 6.18: Processo de Seleção de Variáveis para o modelo preditivo.

iterativa foi executada, agora a partir do modelo de regressão. Uma nova seleção dos atributos foi feita e o critério de menor significância (valor-P) foi adotado para a exclusão.

	Iteração 1		Iteração 2 - Modelo 1		Iteração 3 - Modelo 2	
Variável Independente	**VIF**	**Valor-P**	**VIF**	**Valor-P**	**VIF**	**Valor-P**
taxa_aprovacao_norm	5.070	0.000	4.610	0.000	3.630	0.000
taxa_conclusao_norm	40.550	0.127	4.100	0.017	4.030	0.016
disciplinas_reincidencias	2.390	0.001	1.320	0.006	1.310	0.006
curso_final_bacharelado	3.350	0.486	2.280	0.193	1.710	0.193
ingresso_regular	4.850	0.943	3.310	0.983		
transferencia_interna	1.420	0.282	1.290	0.573	1.280	0.570
formacao_atrasada	1.910	0.736	1.890	0.775	1.830	0.766
disciplinas_aprovado	Infinity	0.006				
disciplinas_reprovado	Infinity	0.01				
disciplinas	Infinity	0.766				

Tabela 6.4: Processo iterativo de seleção de variáveis independentes para iterações 1 a 3. Destaques em vermelho são os indicadores usados para exclusão da variável. Destaques em laranja indicam valores fora de conformidade (VIF > 10 e valor-P > 0.05). Destaques em verde são as variáveis selecionadas ao final (Tabela 6.5).

A Tabela 6.4 representa as primeiras três iterações realizadas para o obtenção do modelo final. Os valores VIF e P são identificados como fora de conformidade como a cor laranja, e quando estão com a cor vermelha são usados como critério de exclusão. Na primeira iteração foram excluídas as variáveis com valor muito alto de multicolinearidade (VIF = infinity) e já a partir da segunda iteração, o valor-P foi usado como referência, seguindo a ordem de menor significância (valor-P > 0.05 e maior valor-P do conjunto). Também a partir da segunda iteração as demais métricas de qualidade começaram a

ser avaliadas. As iterações seguintes, de número 4 a 6, estão descritas na Tabela 6.5, incluindo a relação final de variáveis selecionadas para o modelo (verde). Na Tabela 6.6, as métricas acompanhadas podem ser vistas progredindo (destaque em verde) a cada nova versão do modelo, até chegarem aos seus melhores índices (Modelo 5).

	Iteração 4 - Modelo 3		Iteração 5 - Modelo 4		Iteração 6 - Modelo 5	
Variável Independente	**VIF**	**Valor-P**	**VIF**	**Valor-P**	**VIF**	**Valor-P**
taxa_aprovacao_norm	3.610	0.000	3.550	0.000	2.840	0.000
taxa_conclusao_norm	3.070	0.005	2.940	0.003	2.550	0.005
disciplinas_reincidencias	1.250	0.004	1.250	0.004	1.230	0.004
curso_final_bacharelado	1.700	0.200	1.660	0.236		
ingresso_regular						
transferencia_interna	1.240	0.607				
formacao_atrasada						
disciplinas_aprovado						
disciplinas_reprovado						
disciplinas						

Tabela 6.5: Processo iterativo de seleção de variáveis independentes para iterações 4 a 6.

	Modelo 1		Modelo 2		Modelo 2		Modelo 4		Modelo 5	
Métrica / Conj. de dados	**Treino**	**Teste**	**Treino**	**Teste**	**Treino**	**Teste**	**Treino**	**Teste**	**Treino**	**Teste**
Acurácia	0.84	0.78	0.84	0.78	0.84	0.79	0.84	0.8	0.84	0.81
Precisão	0.9	0.85	0.91	0.85	0.91	0.85	0.91	0.87	0.91	0.87
Revocação	0.84	0.81	0.84	0.81	0.84	0.82	0.84	0.82	0.84	0.83
F1	0.87	0.83	0.87	0.83	0.87	0.83	0.87	0.84	0.87	0.85
ROC-AUC	0.84	0.77	0.84	0.77	0.84	0.78	0.84	0.79	0.84	0.8

Tabela 6.6: Progressão dos valores das métricas de qualidade das diversas versões de modelos testadas. O destaque em verde representa os valores melhorados.

Em uma análise inicial e parcial, a partir dos valores resultantes do teste de *Pearson* (Tabela 6.7), se percebe que a "taxa de aprovação" (taxa_aprovacao_norm) e a "taxa de conclusão" (taxa_conclusao_norm) são as variáveis que mais se correlacionam[12] com a variável dependente "evadiu". Isto é, quanto maior o aproveitamento que o alunos faz das disciplinas que cursa, menor a sua probabilidade de evasão. Esta mesma lógica se aplica em relação ao seu progresso no curso. A indicação de "disciplinas reincidentes" (disciplinas_reincidencias), pode ser interpretada livremente, como um fator de demons-

[12] O sinal negativo do coeficiente de correlação indica que elas se relacionam de forma contrária, ou seja, quanto maior o valor X, maior a relação com o inverso de Y.

tração do comprometimento do aluno em concluir sua formação. Quanto maior o número de reincidências, menor sua chance de desistir do curso, segundo o modelo.

Variável Independente	**Pearson**
taxa_aprovacao_norm	-0.604
taxa_conclusao_norm	-0.515
disciplinas_reincidencias	-0.143
evadiu	1.000

Tabela 6.7: Fator de correlação de *Pearson* para as variáveis independentes usadas no modelo e a variável dependente (evadiu).

7. ANÁLISE DO MODELO

Neste capítulo, são apresentados os resultados obtidos através da análise da implementação do modelo proposto. O objetivo da análise foi verificar se, com os requisitos contemplados, a implementação do modelo poderia auxiliar os especialistas a acompanharem o percurso acadêmico dos estudantes. Este Capítulo, além da descrição da metodologia e perfil dos participantes (Seção 7.1), apresenta os resultados destas entrevistas na Seção 7.2. Na Seção 7.3, são apresentados alguns estudos de caso, mostrando como a implementação baseada no modelo de visualização pode auxiliar nas análises de cenários propostos. Por fim, na Seção 7.4, são apresentadas suas contribuições e limitações.

7.1 Metodologia e Perfil dos Participantes

A metodologia adotada no processo de análise da implementação foi a de entrevistas semiestruturadas com questões abertas em profundidade, seguindo o mesmo método adotado nas entrevistas com especialistas descrito no Capítulo 5. Detalhes do roteiro e questionário aplicado, que foi previamente aprovado pelo Comitê de Ética em Pesquisa por meio de uma emenda, são apresentado no Apêndice B.

Inicialmente, foram convidados os quatro participantes da primeira entrevista, mas dois deles não puderam participar. Portanto, foram convidados outros dois novos participantes, com o mesmo perfil de experiência nas áreas relacionadas ao acompanhamento de alunos no ensino superior. A inclusão de novos participantes também se mostrou importante para se ter um retorno sobre o modelo proposto sem o viés de quem participou do levantamento de requisitos. As entrevistas foram feitas com três homens e uma mulher, com perfil de escolaridade de pós-graduados (doutores e mestres), e atuando em instituições de ensino privadas e públicas. O tempo de experiência variou de 5 a 15 anos, nas funções de gestão, coordenação ou de análise de dados educacionais.

A entrevista foi dividida em quatro momentos. Inicialmente, foram introduzidas as premissas da pesquisa e foi feito um rápido resumo do histórico de desenvolvimento do modelo (nesta etapa alguns novos perfis foram coletados). Na sequência, a implementação do modelo visual foi apresentada por meio de uma breve explicação da sua organização (acompanhamento, detalhamento gráfico e modelo preditivo). Após, foi solicitado que os especialistas explorassem a implementação, para que tivessem suas próprias impressões e, por fim, foram feitas algumas perguntas a respeito destas impressões e sobre os cenários de utilização do modelo para o acompanhamento de alunos. As entrevistas foram realizadas de forma *online* e duraram cerca de 60 minutos, com comentários e questionamentos feitos a qualquer momento de forma espontânea e informal. Todas as entrevistas foram gravadas para posterior transcrição e análise.

7.2 Análise das Entrevistas

As entrevistas aconteceram de forma informal e com algumas breves interrupções, visto que três dos participantes ainda está exercendo suas atividades em regime de teletrabalho. Após feitas as apresentações iniciais e levantamento de perfil da primeira etapa do processo de entrevistas (os dois novos entrevistados estavam tendo o primeiro contato com este estudo), a implementação do modelo foi apresentada e logo começaram as surgir os primeiros comentários e questionamentos.

A visualização de percurso se mostrou de fácil entendimento, juntamente com as movimentações de fluxo de entrada e saída de semestres. O seguinte comentário de um dos participantes valida esta afirmação: *"A visão é boa, eu vejo as entradas [do semestre], as evasões ..."*. Alguns participantes já começaram a solicitar algumas análises no momento da apresentação do modelo, por exemplo: *"Seleciona um período menor [via filtros]. Vamos ver como foi a movimentação."*. O diagrama *Sankey* foi descrito como *"Super interessante e fácil de entender"*, por um dos participantes. A funcionalidade de navegação panorâmica e ampliação centrada foram consideradas *"muito boas"*. Foi mencionado que poderia ser interessante ter uma forma de sinalizar o final do período de dados no diagrama de percurso.

Considerando as demais visualizações interativas, tais como sumário, detalhamento de alunos e disciplinas, foi possível observar os participantes navegando e interpretando os gráficos (em um dos casos o participante somente pôde narrar como estava interagindo). Por exemplo, ao clicar em uma ligação vinda de uma disciplina em direção ao agrupamento de Reprovados, o componente de Detalhamento de Alunos foi filtrado e o participante comentou: *"vejo que tu tens dois alunos neste grupo ..."*. Foi sugerido que as células das visões tabulares fossem colorizadas de acordo com o seus valores, como uma visão de mapa de calor. Um participante comentou que alguns rótulos das colunas poderiam ser mais claros, e sobre a falta de uma documentação geral. A funcionalidade de interação por *crossfilter* foi elogiada.

Sobre as opções do diagrama foram feitos alguns comentários, como por exemplo, que a seleção do esquema de cores poderia ser feita simultaneamente para as diversas seções e não apenas de forma individual. A funcionalidade de realce de cores usando a taxa de aprovação como referência nas disciplinas recebeu o comentário: *"[o realce de cor por disciplina] tem grande utilidade"*. Sugestões foram feitas sobre as cores de sinalização utilizadas no sentido de serem configuráveis pelo usuário e estarem presentes em uma legenda para facilitar sua interpretação.

Ao longo da entrevista algumas dúvidas surgiram, como: Uma dúvida de um participante ao longo da entrevista era por que o nó representando *Aprovados* era maior que a ligação conectada a ele no final de semestre. Após a explicação (descrito na Seção 6.3),

o participante ponderou se este comportamento não poderia levar a uma interpretação equivocada. Por isso, ele sugeriu a opção de colapsar os nós de disciplinas para que fossem vistos alunos únicos, frisando ter entendido o conceito e as limitações de visualização de múltiplos momentos do mesmo aluno em um único semestre. Estas considerações, que certamente são importantes para o domínio deste entrevistado, não foram mencionadas pelos demais.

Um participante fez a seguinte consideração sobre a movimentação temporal: *"... quando um aluno repete uma disciplina ele fica preso ao semestre anterior"*, e questionou se este comportamento poderia ser visualizado. Mas, em seguida ele mesmo ponderou sobre a real necessidade desta funcionalidade, já que é possível filtrar por usuários de acordo com o número de reincidências e realizar o acompanhamento destas ao longo do tempo.

Em relação aos dados visualizados pelo diagrama, outro questionamento foi sobre a possibilidade de se trazerem detalhamentos sobre outras formas de ingresso na visão, como por exemplo, um novo bloco representando Reingressos ou subgrupos detalhando outras informações do alunos (ex: SiSU, ProUni e FIES)[1].

A área de detalhamento gráfico, quando apresentada foi considerada interessante e útil por três dos participantes. Um participante comentou: *""Estou surpreso com a quantidade de informação e com a boa escolha da organização [das informações]""*. Outro mencionou a importância da possibilidade de exploração e se poderiam ter outros formatos de gráficos, mas não especificou quais. Um dos participantes se mostrou indiferente sobre este recurso e não teceu comentários a respeito.

O modelo estatístico de predição foi apresentado e recebeu comentários positivos:

- *"hoje temos alternativas para a visualização [visões de acompanhamento], mas não temos nada comparado ao modelo criado [e conectado a visualização de percurso]"*
- *"A visualização é sobre o passado ... e tem sua relevância, o modelo me dá a chance de prever o futuro..."*
- *"Com certeza vai ser muito útil."*
- *"Hoje nós conseguimos viver sem a visualização ... embora seja importante ... mas em questão de modelos ... é um buraco maior para nós."*

O predição de evasão foi considerada de *"grande valor"*, ao modelo proposto. Já a interface do modelo preditivo foi considerada *"muito interessante"*, por possibilitar a alteração de parâmetros do modelo preditivo em tempo real e foi sugerido que uma documentação

[1] Programas do governo que permitem, respectivamente: ingresso em instituições públicas por pré-qualificação, bolsas integrais ou parciais e acesso a financiamento facilitado.

mais robusta seja criada para está seção. Uma questão sobre o modelo preditivo foi feita a respeito de como são usados os dados de cada usuário ao longo do período avaliado (se o modelo usa todo o conjunto disponível ou somente o período visualizado/filtrado). Foi explicado que o modelo é treinado com todo o conjunto de dados, usando a razão padrão para o treino de 70%, porém os dados considerados para o cálculo de probabilidade de evasão dependem dos valores dos atributos em um determinado período. Isto leva a uma situação na qual, quanto maior o período de dados disponível para a predição, maior são as possibilidades de uma predição com maior acurácia. Um participante comentou que as variáveis independentes usadas para o modelo (taxa de aprovação, taxa de conclusão e número de reincidências) faziam *"sentido"*. Ele também ponderou que o número de reincidências é as vezes considerado um fator de evasão, pelo senso comum. Outro participante considerou o modelo preditivo um componente de grande contribuição para o trabalho, complementando: *"isso que foi criado é algo independente [modelo preditivo], que poderia ser facilmente conectado internamente"*.

As percepções gerais foram muito positivas no sentido que a implementação do modelo é vista como algo de valor por todos os entrevistados e que pode contribuir para o acompanhamento de estudantes, conformem mostram os seguintes comentários: *"... estou muito positivamente surpreso [com a implementação do modelo]"*, *"... [a implementação do modelo] está perto de ser um produto"*, ou do tipo *"... como podemos usar isso hoje?"*. Este último comentário também ajuda a entender realidade de alguns dos entrevistados: *"[a respeito do acompanhamento de estudantes] estamos desatendidos, embora tenhamos uma ferramenta de BI (Business Inteligence)"*.

Ainda sobre a implementação, um entrevistado comentou: *"a forma de visualização é boa e está adequada"*, mas ponderou que existem outras informações usadas na instituição e recomendou que novas rodadas de iteração fossem feitas com usuários para que o modelo fique ainda mais alinhado com a necessidades do dia-a-dia. Pois segundo sua experiências, duas iterações não seriam suficientes para que todas as funcionalidades fossem completamente mapeadas e implementadas. Os demais participantes não fizeram comentários neste sentido.

Dois participantes fizeram comentários a respeito da performance da implementação. Alertando para o grande volume de dados das suas instituições em caso de uma instalação em ambiente real.

Ao longo desta análise de entrevistas, algumas sugestões de melhorias foram trazidas por meio de comentários dos participantes. Estas sugestões foram compiladas e são apresentadas na próxima seção.

7.3 Estudo de Casos

Nesta seção são apresentados alguns dos possíveis cenários de uso da implementação do modelo visual, e como ela pode auxiliar na análise dos dados.

1. **Identificação de formação em atraso:** A identificação do número de semestres cursados e se o aluno está com a formação em atraso é algo relevante, pois conforme o perfil da instituição, pode representar um potencial problema na formação do estudante (podendo inclusive ser justificada por questões socioeconômico [28]). Na implementação, conforme ilustra a Figura 7.1, estes indicadores podem ser acompanhados a partir do *Detalhamento (Aluno/Disciplina)*, nas colunas *total_semestres* e *formacao_atrasada* (Figura 7.1A) (este último indicador pode ser encontrado também, como um filtro lateral (Figura 7.1B)).

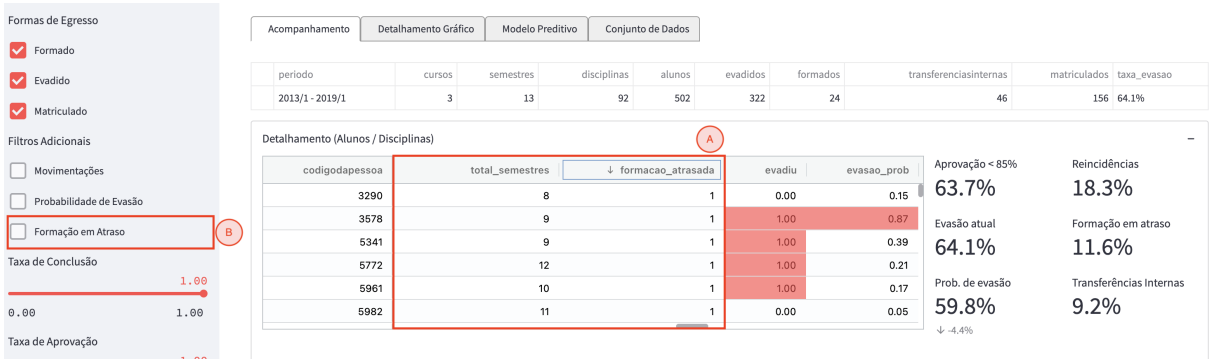


Figura 7.1: Estudo de caso 1: Identificação de formação em atraso.

2. **Probabilidade de evasão de alunos em uma turma:** Entendendo que a evasão é algo a ser evitado nas instituições, o monitoramento da probabilidade de evasão dos alunos pode ser usado para direcionar ações que efetivamente evitem esta forma de egresso. Pela Figura 7.2, a partir do modelo preditivo disponível, é possível levantar este indicador para cada alunos em uma turma (Figura 7.2A) (ou outros agrupamentos de alunos), selecionando no Filtro, a caixa de seleção - *Probabilidade de Evasão* (Figura 7.2B). Os detalhes da métrica (*evasao_prob*) serão filtrados e disponibilizados na visão *Detalhamento (Aluno/Disciplina)* (Figura 7.2C).

3. **Disciplinas com maior índice de reprovação:** A reprovação em disciplinas, a partir das análises realizadas, é um dos motivadores da evasão. Portanto, entender quais disciplinas apresentam alto índice de reprovação pode ajudar em direcionamentos que melhorem o desempenho dos alunos nas mesmas. A Figura 7.3 demonstra como o acompanhamento destas disciplinas pode ser feito através da visualização de *Percurso Acadêmico*. Utilizando a opção de visualização de *Usar realce por cores* (Figura 7.3A), seguida pela caixa de seleção de *Taxa de aprovação* (Figura

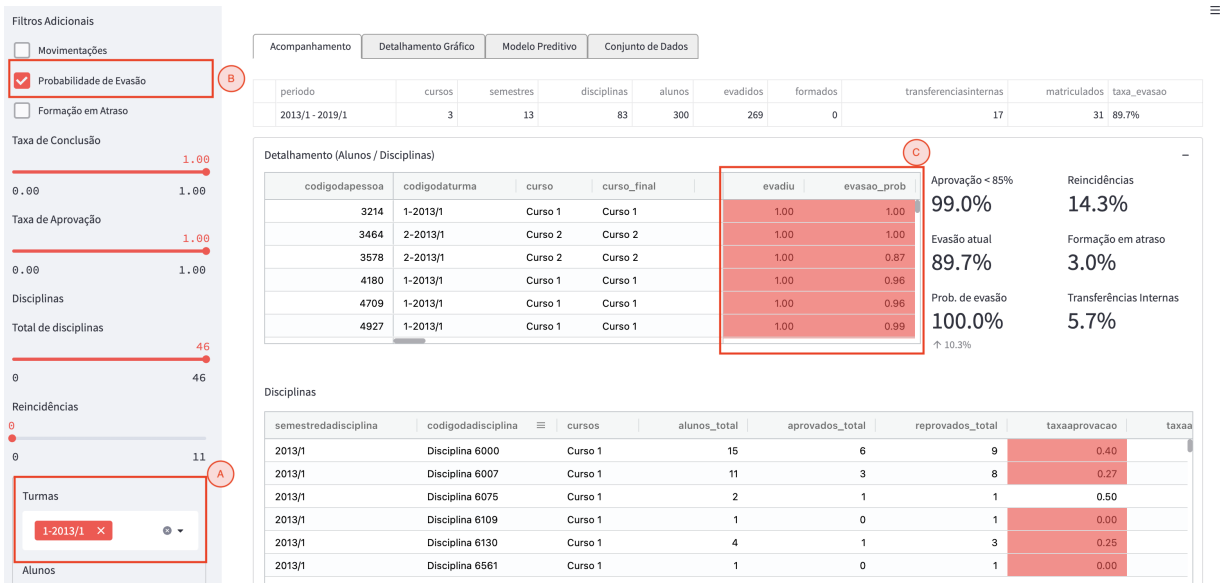


Figura 7.2: Estudo de caso 2: Probabilidade de evasão de uma turma.

7.3B), a visão de percurso representará as disciplinas com maior índice de reprovação destacadas em vermelho (Figura 7.3C), facilitando assim, sua identificação.

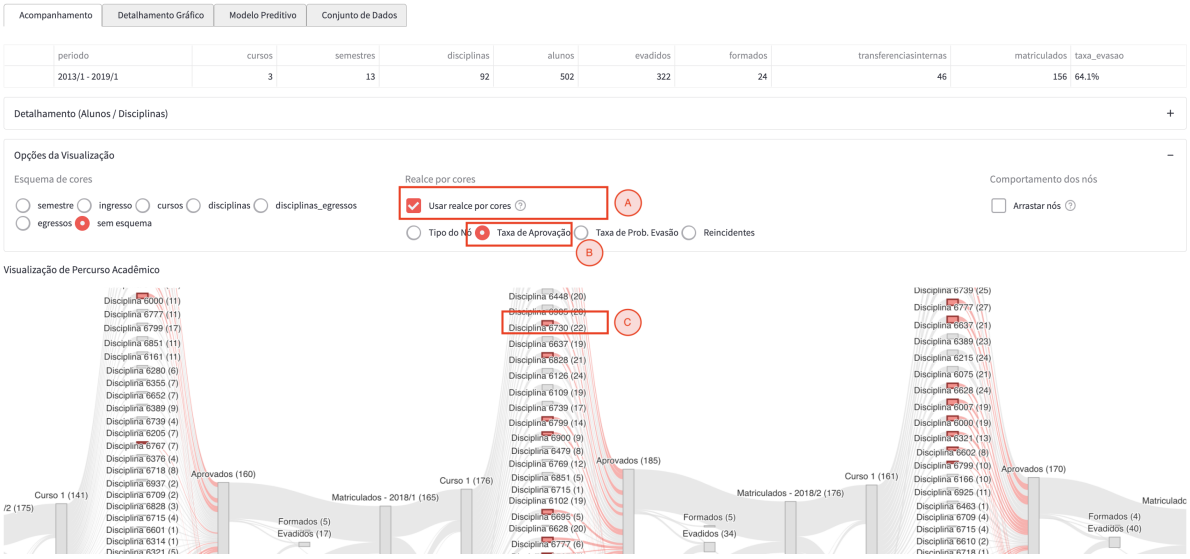


Figura 7.3: Estudo de caso 3: Disciplinas com maior índice de reprovação.

7.4 Contribuições e Limitações

As contribuições deste trabalho foram, primeiramente, a revisão sistemática da literatura sobre análise visual do percurso acadêmico ao longo do ensino superior. Devido a especificidade do tema, poucos trabalhos foram encontrados, desta forma, espera-se que a RSL apresentada, que inclui a descrição dos trabalhos e uma discussão sobre as

técnicas de visualização utilizadas e suas contribuições, possa servir de base para futuras pesquisas nesta área. Em segundo lugar, o conjunto de requisitos identificados através da RSL e das entrevistas com especialistas de domínio, pode auxiliar no projeto e desenvolvimento de outros trabalhos relacionados à análise de dados educacionais. O modelo para análise visual do percurso acadêmico de estudantes e a implementação deste, através de uma visualização interativa baseada em diagrama de *Sankey*, conectada a um modelo preditivo de evasão, também são contribuições deste trabalho. Juntamente com as interfaces interativas do modelo preditivo e de detalhamento gráfico, usada para análise exploratória de dados, que podem ser utilizadas a partir da implementação e eventualmente conectadas a outros sistemas nas instituições, conforme mencionado nas entrevistas. O código da implementação do modelo está disponível de forma aberta, atendendo à R13 [2].

De forma geral, a implementação do modelo visual, segundo os próprios participantes entrevistados e como exemplificado pelos cenários descritos na Seção 7.3, poderia complementar o conjunto de ferramentas usadas nas instituições, e assim contribuir para a melhora do acompanhamento estudantil.

Quanto a implementação proposta, a Tabela 7.1 apresenta algumas das sugestões de melhorias citadas nas entrevistas de análise do modelo implementado, e que foram discutidas na Seção 7.2.

Além das sugestões citadas, houveram pedidos para que sejam feitos testes e ajustes para manter o bom desempenho da implementação a partir do uso de um grande conjunto de dados (cenário das instituições de ensino). Outra solicitação, foi em relação a necessidade de serem incorporadas outras informações acadêmicas ao perfil do aluno. Neste sentido, novas rodadas de iteração com usuários foram sugeridas, para que sejam adicionadas novas dimensões e métricas ao modelo.

Outras limitações identificadas neste trabalho são apresentadas a seguir:

- Generalizar a implementação do modelo visual: limitada inicialmente pelo conjunto de dados disponível, e posteriormente adotada com a intenção de tornar o modelo facilmente adaptável às instituições de ensino, pode não atender a todos os casos, conforme mencionado nas entrevistas de análise. Por isso, seria necessário trabalhar na sua generalização.

- Incluir outros modelos de predição: considerando outros possíveis cenários e diferentes dados, talvez o modelo predição adotado, usando regressão logística, não seja o mais indicado, sendo necessário a implementação e teste de outras técnicas.

- Permitir o acompanhamento da ordem de realização das disciplinas (R9): embora este requisito tenha sido atendido no modelo, poderia ser melhor explorado através de técnicas de mineração de dados sequenciais.

[2] https://github.com/DAVINTLAB/academic-path-visualization

Área	Descrição
Acompanhamento	• Utilizar padrão de cores no estilo mapa de calor, para os valores nas células das visões tabulares. • Sinalizar o final do diagrama de percurso; • A seleção do esquema de cores poder ser feita simultaneamente para as diversas seções; • Cores de sinalização utilizadas poderiam ser configuráveis pelo usuário; • Legenda para a interpretação das cores de sinalização no diagrama de percurso; • Opção de colapsar os nós de disciplinas para que possam ser vistos alunos únicos de forma agrupada; • Possibilidade de se trazerem detalhamentos sobre as formas de ingresso. • Possibilidade configurar o tamanho da fonte e comportamento dos rótulos de descrição dos nós no diagrama. • Documentação e melhores descrições dos rótulos utilizados.
Detalhamento gráfico	• Possibilidade de outros arranjos visuais além dos gráficos de barras empilhados e histogramas.
Modelo preditivo	• Documentação mais detalhada para a seção do modelo preditivo.

Tabela 7.1: Sugestões de melhorias identificadas pelas entrevistas de análise da implementação do modelo visual.

- Possibilitar a inclusão de outros conjuntos de dados no modelo visual: o acompanhamento estudantil e, em especial, a evasão, são problemas complexos e que requerem outras análises além das apresentadas neste trabalho. Por exemplo, o uso de dados socioeconômicos pode ter correlação com o desempenho do aluno. Desta forma, este trabalho, não se propõe a realizar uma análise em profundidade das motivações que levam um aluno apresentar um baixo desempenho ou evadir, mas pode ser estendido para isso.

- Trabalhar na melhoria de algumas limitações técnicas: testar e validar a conexão de dados com as instituições de ensino superior, pois é suportado internamente pela implementação, mas precisa ser testado com os bancos relacionais utilizados; usar mais informações para os detalhamentos (o que é suportado via parametrização

interna); e melhorar o desempenho da implementação, o que necessita um trabalho adicional de otimização.

8. CONCLUSÃO E TRABALHOS FUTUROS

Este trabalho apresentou um modelo para análise visual do percurso acadêmico de estudantes, que visa ajudar os tomadores de decisão das instituições de ensino superior, no acompanhamento de alunos e grupos de alunos com risco de não concluírem seus cursos com sucesso. Este modelo foi desenvolvido a partir de um conjunto de requisitos identificados através de uma revisão da literatura e de entrevistas com quatro especialistas de domínio, para entender suas realidades e necessidades. A RSL também permitiu identificar uma lacuna em relação ao uso de visualizações para acompanhamento do progresso de estudantes, associadas a modelos de predição de evasão. O resultado desta pesquisa, então, culminou na proposição deste modelo que é centrado em uma visualização que utiliza diagrama de *Sankey*, e está conectado a um modelo de predição de evasão baseado em regressão logística.

A partir dos relatos dos especialistas de domínio, coletados por meio de entrevistas de análise deste estudo, constatou-se que o uso de diagrama de *Sankey* junto com os demais elementos visuais do modelo, mostrou-se promissor para ajudar na exploração e análise visual de dados dos estudantes. O modelo preditivo também foi elogiado, pois o indicador de probabilidade de evasão pode ajudar a orientar ações que efetivamente evitem esta situação. A implementação do modelo contempla os requisitos levantados e está alinhada com a questão de pesquisa (QP), pois através das visualizações e interações fornecidas, permite tanto uma visão ampla do percurso acadêmico, quanto um maior detalhamento de indivíduos quando necessário, auxiliando, assim, no acompanhamento do percurso acadêmico dos estudantes. Desta forma, pode contribuir para que os agentes responsáveis possam tomar ações em prol do sucesso dos estudantes utilizando uma abordagem baseada em evidências e orientada por dados (em contrapartida a uma abordagem baseada em experiências e orientada por ideias [53]).

Portanto, as principais contribuições deste trabalho são a RSL, o conjunto de requisitos e oportunidades que podem auxiliar na análise do percurso acadêmico dos estudantes, o modelo visual proposto e a sua implementação.

As limitações estão detalhadas na Seção 7.4, juntamente com as melhorias propostas pelos especialistas de domínio entrevistados. Considerando as limitações apresentadas, inicialmente pretende-se fazer as melhorias propostas pelos especialistas na implementação do modelo.

Como trabalhos futuros, ampliando os cenários contemplados, será feito um estudo para alterar o modelo visual e permitir a inclusão de outros tipos de dados, tais como dados socioeconômicos ou relacionados ao contexto educacional durante (e pós) a COVID-19. Também será explorada a inclusão de outros modelos preditivos, buscando alcançar melhores resultados. Por fim, espera-se fazer novas análises do modelo visual

proposto e poder aplicá-lo em um ambiente real, a partir da simulação de novos cenários de acompanhamento.

Finalizando, espera-se que este trabalho possa contribuir para a pesquisa relacionada ao acompanhamento do percurso acadêmico de estudantes ao longo do ensino superior, e que possa ser utilizado como ferramenta de auxílio às análises feitas em diferentes instituições de ensino.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] Aggarwal, V.; Kosian, S. “Feature selection and dimension reduction techniques in sas”, *EXL Service*, vol. 01, Sep 2011, pp. 6.

[2] Alban, M.; Mauricio, D. “Predicting university dropout through data mining: A systematic literature”, *Indian Journal of Science and Technology*, vol. 12–4, Feb 2019, pp. 12.

[3] Aldowah, H.; Al-Samarraie, H.; Fauzy, W. M. “Educational data mining and learning analytics for 21st century higher education: A review and synthesis”, *Telematics and Informatics*, vol. 37, Apr 2019, pp. 13–49.

[4] American Academy of Arts & Sciences. “A primer on the college student journey”. Capturado em: https://www.amacad.org/publication/primer-college-student-journey, 2022-05-27.

[5] Askinadze, A.; Liebeck, M.; Conrad, S.; Heine, H. “Using venn, sankey, and upset diagrams to visualize students’ study progress based on exam combinations”. In: Proceedings of the 9th International Conference on Learning Analytics and Knowledge, 2019, pp. 1–5.

[6] Baggi, C. A. D. S.; Lopes, D. A. “Evasão e avaliação institucional no ensino superior: uma discussão bibliográfica”, *Avaliação: Revista da Avaliação da Educação Superior*, vol. 16–2, Jul 2011, pp. 355–374.

[7] Bakken, S. S.; Suraski, Z.; Schmid, E. “PHP Manual: Volume 1”. iUniverse, Incorporated, 2000, 480p.

[8] Basavaraj, P.; Badillo-Urquiola, K.; Garibay, I.; Wisniewski, P. J. “A tale of two majors: When information technology is embedded within a department of computer science”. In: Proceedings of the 19th Annual SIG Conference on Information Technology Education, 2018, pp. 32–37.

[9] Booth, A.; Sutton, A.; Clowes, M.; Martyn-St James, M. “Systematic approaches to a successful literature review”. SAGE Publications Ltd, 2021, 424p.

[10] Bostock, M.; Ogievetsky, V.; Heer, J. “D^3 data-driven documents”, *IEEE transactions on visualization and computer graphics*, vol. 17–12, Dec 2011, pp. 2301–2309.

[11] Brasil. “Lei nº 13.709, de 14 de agosto de 2018”. Capturado em: http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13709compilado.htm, 2022-05-15.

[12] Cabello, A. F.; Falqueto, J.; Zandonade, M.; Ferreira, G. V.; Arruda, J. A. D.; Alvarez, G. A.; Imbroisi, D. “Evasão no ensino superior: qual metodologia adotar? uma análise sobre o efeito de diferentes metodologias para a identificação dos índices de evasão no ensino superior brasileiro”. In: Proceedings of the Colóquio Internacional De Gestão Universitária, 2018, pp. 14.

[13] Campbell, J. P.; DeBlois, P. B.; Oblinger, D. G. “Academic analytics: A new tool for a new era”, *EDUCAUSE review*, vol. 42–4, Jul 2007, pp. 40–57.

[14] Card, M. “Readings in information visualization: using vision to think”. Morgan Kaufmann, 1999, 686p.

[15] Charleer, S.; Klerkx, J.; Duval, E. “Learning dashboards”, *Journal of Learning Analytics*, vol. 1–3, Dec 2014, pp. 199–202.

[16] Chen, C.-h.; Härdle, W.; Unwin, A. “Handbook of Data Visualization”. Springer, 2008, 936p.

[17] Chok, N. S. “Pearson’s versus spearman’s and kendall’s correlation coefficients for continuous data”, Tese de Doutorado, University of Pittsburgh, 2010, 53p.

[18] Cook, K. A.; Thomas, J. J. “Illuminating the path: The research and development agenda for visual analytics”. National Visualization and Analytics Ctr, 2005, 190p.

[19] Coussement, K.; Phan, M.; De Caigny, A.; Benoit, D. F.; Raes, A. “Predicting student dropout in subscription-based online learning environments: The beneficial impact of the logit leaf model”, *Decision Support Systems*, vol. 135–113325, Aug 2020, pp. 33.

[20] Dermeval, D.; Coelho, J. A. P. d. M.; Bittencourt, I. I. “Mapeamento Sistemático e Revisão Sistemática da Literatura em Informática na Educação”. SBC, 2019, cap. 3, pp. 26.

[21] Domingos, P. “The role of occam’s razor in knowledge discovery”, *Data mining and knowledge discovery*, vol. 3–4, Dec 1999, pp. 409–425.

[22] Felizardo, K. R.; Mendes, E.; Kalinowski, M.; Souza, É. F.; Vijaykumar, N. L. “Using forward snowballing to update systematic reviews in software engineering”. In: Proceedings of the 10th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2016, pp. 6.

[23] Fernandes, A. A. T.; Filho, D. B. F.; da Rocha, E. C.; da Silva Nascimento, W. “Leia este artigo se você quiser aprender regressão logística”, *Revista de Sociologia e Política*, vol. 28–74, Jun 2020, pp. 20.

[24] Ferreira, F.; Santos, B. S.; Marques, B.; Dias, P. “FICAvis: Data Visualization to Prevent University Dropout”. In: Proceedings of the 24th International Conference Information Visualisation (IV), 2020, pp. 57–62.

[25] Ferreira, J. C.; Patino, C. M. “What does the p value really mean?”, *Jornal Brasileiro de Pneumologia*, vol. 41–5, Sep 2015, pp. 485.

[26] Gilbert, J. K.; Boulter, C. “Developing models in science education”. Springer, 2012, 387p.

[27] Greer, J. E.; Thompson, C.; Banow, R.; Frost, S. “Data-Driven Programmatic Change at Universities: What works and how”. In: Proceedings of the PCLA@ LAK, 2016, pp. 32–35.

[28] Gubbala, N.; Baynes, A. “Graduated Life Explore: An Interactive Visualization Tool to Explore the Comprehensive Benefits of a Timely Graduation”. In: Proceedings of the IEEE Frontiers in Education Conference (FIE), 2019, pp. 1–5.

[29] Guyon, I.; Elisseeff, A. “An introduction to variable and feature selection”, *Journal of machine learning research*, vol. 3–03, Mar 2003, pp. 1157–1182.

[30] Hartmann, I. A. “Lgpd e pesquisa acadêmica”. Capturado em: https://bibliotecadigital.fgv.br/dspace/bitstream/handle/10438/27962/10%20-%20Ivar%20Hartman%20-%20LGPD%20e%20Pesquisa%20Acad%C3%AAmica%2021ago19.pdf, 2022-05-15.

[31] Heileman, G. L.; Babbitt, T. H.; Abdallah, C. T. “Visualizing Student Flows: Busting Myths About Student Movement and Success”, *Change: The Magazine of Higher Learning*, vol. 47–3, May 2015, pp. 30–39.

[32] Horvath, D. M.; Molontay, R.; Szabo, M. “Visualizing Student Flows to Track Retention and Graduation Rates”. In: Proceedings of the 22nd International Conference Information Visualisation (IV), 2018, pp. 338–343.

[33] Ihaka, R.; Gentleman, R. “R: a language for data analysis and graphics”, *Journal of computational and graphical statistics*, vol. 5–3, Sep 1996, pp. 299–314.

[34] INEP. “Censo da educação superior 2000-2020”. Capturado em: https://download.inep.gov.br/publicacoes/institucionais/estatisticas_e_indicadores/notas_estatisticas_censo_da_educacao_superior_2020.pdf, 2022-05-10.

[35] Jankun-Kelly, T.; Ma, K.-L.; Gertz, M. “A model and framework for visualization exploration”, *IEEE Transactions on Visualization and Computer Graphics*, vol. 13–2, Apr 2007, pp. 357–369.

[36] Jørnø, R. L.; Gynther, K. “What constitutes an ‘actionable insight’ in learning analytics?”, *Journal of Learning Analytics*, vol. 5–3, Dec 2018, pp. 198–221.

[37] Kantorski, G.; Flores, E. G.; Schmitt, J.; Hoffmann, I.; Barbosa, F. “Predição da evasão em cursos de graduação em instituições públicas”, *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação - SBIE)*, vol. 27–11, Nov 2016, pp. 906.

[38] Kee, D. E.; Salowitz, L.; Chang, R. “Comparing interactive web-based visualization rendering techniques”. In: Proceedings of the IEEE Symposium on Information Visualization (INFOVIS), 2012, pp. 2.

[39] Kitchenham, B.; Charters, S. “Guidelines for performing systematic literature reviews in software engineering”, Relatório Técnico, School of Computer Science of Keele University and Department of Computer Science of Durham University, 2007, 65p.

[40] Klymkowsky, M.; Martin, A.; Stubbs, R.; Oran, A. “Identifying Students’ Progress and Mobility Patterns in Higher Education Through Open-Source Visualization”, *SocArXiv*, vol. 0, Jun 2019, pp. 1–19.

[41] Kwak, C.; Clayton-Matthews, A. “Multinomial logistic regression”, *Nursing research*, vol. 51–6, Nov 2002, pp. 404–410.

[42] Lazar, J.; Feng, J. H.; Hochheiser, H. “Research methods in human-computer interaction”. Morgan Kaufmann, 2017, 560p.

[43] Li, D.; Mei, H.; Shen, Y.; Su, S.; Zhang, W.; Wang, J.; Zu, M.; Chen, W. “Echarts: A declarative framework for rapid construction of web-based visualization”, *Visual Informatics*, vol. 2–2, Jun 2018, pp. 136–146.

[44] Lin, S.; Fortuna, J.; Kulkarni, C.; Stone, M.; Heer, J. “Selecting semantically-resonant colors for data visualization”. In: Proceedings of the 15th Eurographics Conference on Visualization, 2013, pp. 401–410.

[45] Longhurst, R. “Semi-structured interviews and focus groups”, *Key methods in geography*, vol. 3–2, Jan 2003, pp. 143–156.

[46] Melançon, G.; Herman, I. “Dag drawing from an information visualization perspective”. In: Proceedings of the Joint Eurographics and IEEE VGTC Symposium on Visualization, 2000, pp. 3–12.

[47] Mowery, K.; Shacham, H. “Pixel perfect: Fingerprinting canvas in html5”. In: Proceedings of the IEEE Workshop on Web 2.0 Security and Privacy (W2SP), 2012, pp. 12.

[48] Neale, P.; Boyce, C. “A guide for designing and conducting in-depth interviews for evaluation input”, *Pathfinder Internatıonal*, vol. 2, May 2006, pp. 16.

[49] Owens, M. “The definitive guide to SQLite”. Apress, 2006, 370p.

[50] O’Handley, B. J.; Ludwig, M. K.; Allison, S. R.; Niemier, M. T.; Kumar, S.; Bualuan, R.; Wang, C. “CoursePathVis: Course Path Visualization Using Flexible Grouping and Funnel-Augmented Sankey Diagram”, *Electronic Imaging*, vol. 34, Jan 2022, pp. 9.

[51] R Core Team. “R: A language and environment for statistical computing”. Capturado em: http://www.R-project.org/, 2022-05-27.

[52] Rahman, F.; Devanbu, P. “How, and why, process metrics are better”. In: Proceedings of the 35th International Conference on Software Engineering (ICSE), 2013, pp. 432–441.

[53] Raji, M.; Duggan, J.; DeCotes, B.; Huang, J.; Vander Zanden, B. “Visual progression analysis of student records data”. In: Proceedings of the IEEE Visualization in Data Science (VDS), 2017, pp. 31–38.

[54] Ramos, J. L. C.; Gomes, A. S.; Rodrigues, R.; Silva, J.; de Souza, F. d. F.; de Gouveia Zambom, E.; Prado, L. “Um modelo preditivo da evasao dos alunos na ead a partir dos construtos da teoria da distância transacional”. In: Proceedings of the XXVIII Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educaçao - SBIE), 2017, pp. 1227.

[55] Ramos, J. L. C.; Rodrigues, R. L.; Silva, J. C. S.; de Oliveira, P. L. S. “Crisp-edm: uma proposta de adaptação do modelo crisp-dm para mineração de dados educacionais”. In: Proceedings of the XXXI Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educaçao - SBIE), 2020, pp. 1092–1101.

[56] Refaeilzadeh, P.; Tang, L.; Liu, H. “Cross-validation”, *Encyclopedia of database systems*, vol. 5, Jan 2009, pp. 532–538.

[57] Riehmann, P.; Hanfler, M.; Froehlich, B. “Interactive sankey diagrams”. In: Proceedings of the IEEE Symposium on Information Visualization (INFOVIS), 2005, pp. 233–240.

[58] Romero, C.; Ventura, S. “Data mining in education”, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 3–1, Dec 2013, pp. 12–27.

[59] Rosvall, M.; Bergstrom, C. T. “Mapping change in large networks”, *PLOS ONE*, vol. 5–1, Jan 2010, pp. 7.

[60] Scanlan, C. L. “Preparing for the Unanticipated: Challenges in Conducting Semi-Structured, In-Depth Interviews”. SAGE Publications Ltd, 2020, 352p.

[61] Schmidt, M. “The sankey diagram in energy and material flow management”, *Journal of industrial ecology*, vol. 12–2, Apr 2008, pp. 173–185.

[62] Schneider, A.; Hommel, G.; Blettner, M. “Linear regression analysis: part 14 of a series on evaluation of scientific publications”, *Deutsches Ärzteblatt International*, vol. 107–44, Nov 2010, pp. 776–782.

[63] Schreiber-Gregory, D.; Jackson, H.; Bader, K. “Logistic and linear regression assumptions: Violation recognition and control”, *Henry M Jackson Foundation*, vol. 247, Jan 2018, pp. 22.

[64] Senaviratna, N. A. M. R.; Cooray, T. M. J. A.; et al.. “Diagnosing multicollinearity of logistic regression model”, *Asian Journal of Probability and Statistics*, vol. 5–2, Oct 2019, pp. 9.

[65] Shukla, S.; Maheshwari, A.; Johri, P. “Comparative analysis of ml algorithms & stream lit web application”. In: Proceedings of the 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), 2021, pp. 175–180.

[66] Shull, F.; Singer, J.; Sjøberg, D. I. K. “Guide to advanced empirical software engineering”. Springer, 2007, 394p.

[67] Siemens, G.; Long, P. “Penetrating the fog: Analytics in learning and education”, *EDUCAUSE review*, vol. 46–5, Sep 2011, pp. 31–40.

[68] Skurla, C.; O’Neal, D. L. “A Novel Tool to Visualize Student Flow Through the Curriculum”. In: Proceedings of the First-Year Engineering Experience, 2021, pp. 7.

[69] Stearns, B.; Rangel, F.; Firmino, F.; Rangel, F.; Oliveira, J. “Prevendo desempenho dos candidatos do enem através de dados socioeconômicos”. In: Proceedings of the 36th SBC Undergraduate Research Contest, 2017, pp. 2522–2530.

[70] Szklo, M. “Population-based cohort studies”, *Epidemiologic reviews*, vol. 20–1, Mar 1998, pp. 81–90.

[71] Thayer, R. H.; Bailin, S. C.; Dorfman, M. “Software requirements engineerings”. IEEE Computer Society Press, 1997, 2 ed., 600p.

[72] Vaclavek, J.; Kuzilek, J.; Skocilas, J.; Zdrahal, Z.; Fuglik, V. “Learning analytics dashboard analysing first-year engineering students”. In: Proceedings of the European Conference on Technology Enhanced Learning, 2018, pp. 575–578.

[73] Weaver, C. “Multidimensional visual analysis using cross-filtered views”. In: Proceedings of the IEEE Symposium on Visual Analytics Science and Technology, 2008, pp. 163–170.

[74] Wexler, J.; Pushkarna, M.; Bolukbasi, T.; Wattenberg, M.; Viégas, F.; Wilson, J. “The what-if tool: Interactive probing of machine learning models”, *IEEE transactions on visualization and computer graphics*, vol. 26–1, Jan 2020, pp. 56–65.

[75] Witten, I. H.; Frank, E. “Data mining: practical machine learning tools and techniques with java implementations”, *ACM Sigmod Record*, vol. 31–1, Mar 2002, pp. 76–77.

APÊNDICE A – LEVANTAMENTO DE REQUISITOS: QUESTIONÁRIO APLICADO

Instrumento	Roteiro semiestruturado – Questionário
Formato	Roteiro semiestruturado para entrevista presencial ou via videoconferência. Este questionário está dividido em 4 etapas: • Etapa 1: Levantamento de perfil; • Etapa 2: Levantamento de Práticas e Ferramentas, Necessidades dos especialistas e Perfil dos Estudantes; • Etapa 3: Análise da Proposta de Visualização; • Etapa 4: Considerações Finais Com duração prevista entre 1h e 1h e 30 minutos, incluindo a leitura do TCLE e introdução do estudo que está sendo realizado.
Objetivo	Entender e identificar possibilidades a respeito de como as instituições analisam as informações relacionadas ao percurso acadêmico dos estudantes, como forma de melhorar seu desempenho e evitar sua evasão. Neste contexto, percurso acadêmico consiste na representação de toda a movimentação acadêmica feita pelo aluno durante o ciclo de tempo de formação em que faz parte de uma instituição de ensino (disciplinas cursadas, desempenho, mudanças de curso e de instituição, forma de egresso, etc...).
Participantes	Serão recrutados de 3 a 5 voluntários para participarem das entrevistas. Critério de inclusão: indivíduos responsáveis pelas áreas relacionadas ao "acompanhamento do estudante", ou correlatas, isto é, áreas em que são analisados dados de estudantes objetivando seu acompanhamento e boa condução ao longo do curso, em instituições de ensino superior. Não serão considerados indivíduos menores de idade nem pessoas que necessitem de algum tipo de amparo (físico ou cognitivo).

Levantamento de Perfil	• Qual o seu maior grau de escolaridade e em qual área? • Qual seu cargo (função) na instituição? • Há quanto tempo está nesta função (cargo) nesta instituição? • Há quanto tempo está nesta instituição? • Você trabalha com o acompanhamento de alunos e, ou análise de evasão na sua instituição (sucesso do estudante)? Há quanto tempo?
Práticas e Ferramentas, Necessidades dos Especialistas e Perfil dos estudantes	• Como realiza o acompanhamento, de forma geral, dos alunos ao longo do curso? • Quais indicadores são utilizados ou julgados como importantes para ajudar na identificação de alunos com dificuldades e/ou com potencial de evasão? • Quais ferramentas/dados são utilizados para fazer o acompanhamento dos alunos ao longo do curso (notas, disciplinas cursadas, disciplinas canceladas, etc.)? • Utiliza ferramentas que possibilitem visualizar o percurso dos alunos ao longo do curso? Em caso de resposta positiva, quais ferramentas são utilizadas e que tipos de visualizações elas fornecem?

Práticas e Ferramentas, Necessidades dos Especialistas e Perfil dos estudantes	• Quais funcionalidades acha interessante que uma ferramenta deste tipo disponibilize, ou o que sente falta nas ferramentas que utiliza? • É feita alguma análise de dados que mostre uma relação entre alunos com um tempo maior de permanência na instituição (ou seja, ficam além do tempo previsto) com alunos que evadem ou tem mais chances de evasão? • Como são ofertadas as disciplinas que serão ministradas em cada semestre (por exemplo: existem disciplinas que não são oferecidas todos os semestres)? • Os alunos recebem orientação para se inscreverem nas disciplinas disponíveis em cada semestre? • É feita alguma análise de dados que relacione a ordem em que as disciplinas são cursadas a uma possível evasão futura? • A instituição utiliza ferramentas ou dados que permitam verificar com que frequência os alunos mudam de curso internamente? Em caso afirmativo. Saberia com que frequência, alunos mudam de curso internamente? • Existe alguma análise a partir de dados, que relacione uma mudança interna de curso, com uma possível evasão futura? • Contando com a possibilidade que alunos que saíram da instituição possam voltar mais tarde para a mesma, você possui ferramentas ou dados que permitam verificar com que frequência os alunos voltam a ingressar na instituição após evadirem?
Práticas e Ferramentas, Necessidades dos Especialistas e Perfil dos estudantes	• De forma geral, consegue identificar antecipadamente alunos com tendência a abandonarem o curso? Em caso de resposta positiva, como realiza esse controle?

Análise da Proposta de Visualização	A partir do modelo de visualização proposto, com dados anonimizados de 3 cursos de TI, de 2013/1 a 2019/1, responda às seguintes questões: • Quais são suas percepções iniciais? • Consegue identificar claramente as dimensões e métricas apresentadas? • Consegue perceber a movimentação de alunos pelas disciplinas e semestres? • Consegue identificar a movimentação de alunos entre os cursos? • Você gostaria de usar uma ferramenta deste tipo? • Você teria sugestões de funcionalidades a serem incluídas nesta ferramenta?
Considerações Finais	• Acha relevante o tema deste trabalho para o seu contexto institucional? • Podemos contatá-la(o) posteriormente para uma apresentação do modelo final resultante desta pesquisa? • Alguma consideração adicional?

APÊNDICE B – ROTEIRO DE ENTREVISTA PARA ANÁLISE DO MODELO PROPOSTO

Instrumento	Roteiro semiestruturado – Questionário
Formato	Roteiro semiestruturado para entrevista presencial ou via videoconferência. Este questionário está dividido em 4 etapas: • Etapa 1: Apresentação e Levantamento de perfil; • Etapa 2: Análise da Proposta de Visualização; • Etapa 3: Considerações Finais Com duração prevista entre 1h e 1h e 30 minutos, incluindo a leitura do TCLE e introdução do estudo que está sendo realizado.
Objetivo	Avaliar o modelo proposto para análise das informações relacionadas ao percurso acadêmico dos estudantes, como forma de melhorar seu desempenho e evitar sua evasão. Neste contexto, percurso acadêmico consiste na representação de toda a movimentação acadêmica feita pelo aluno durante o ciclo de tempo de formação em que faz parte de uma instituição de ensino (disciplinas cursadas, desempenho, mudanças de curso e de instituição, forma de egresso, etc...).
Participantes	Serão recrutados de 3 a 6 voluntários para participarem das entrevistas. Critério de inclusão: indivíduos responsáveis pelas áreas relacionadas ao "acompanhamento do estudante", ou correlatas, isto é, áreas em que são analisados dados de estudantes objetivando seu acompanhamento e boa condução ao longo do curso, em instituições de ensino superior. Não serão considerados indivíduos menores de idade nem pessoas que necessitem de algum tipo de amparo (físico ou cognitivo).

Levantamento de Perfil	• Qual o seu maior grau de escolaridade e em qual área? • Qual seu cargo (função) na instituição? • Há quanto tempo está nesta função (cargo) nesta instituição? • Há quanto tempo está nesta instituição? • Você trabalha com o acompanhamento de alunos e, ou análise de evasão na sua instituição (sucesso do estudante)? Há quanto tempo?
Análise da Proposta de Visualização	A partir do modelo de visualização proposto, com dados anonimizados de 3 cursos de TI, de 2013/1 a 2019/1, responda às seguintes questões: • Quais são suas percepções iniciais? • Consegue identificar claramente as dimensões e métricas apresentadas? • As visualizações/funcionalidades presentes permitem identificar de maneira eficiente a movimentação dos alunos pelas disciplinas e semestres? Justifique sua resposta. • Consegue identificar a movimentação de alunos entre os cursos? • Você achou o modelo fácil de usar? Quais funcionalidades apresentam maior e menor dificuldade? Justifique sua resposta. • Quais as funcionalidades que você mais gostou? Por quê? • Você gostaria de usar uma ferramenta deste tipo? Por quê? • Você teria sugestões de outras funcionalidades a serem incluídas nesta ferramenta?

Considerações Finais	• Acha relevante o tema deste trabalho para o seu contexto institucional? • Quais sugestões você teria para aprimorar este modelo? Por exemplo, outros indicadores, visualizações ou funcionalidades. • Alguma consideração adicional?

APÊNDICE C – TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO

TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO (TCLE)

Nós, Fernando Lunardelli (aluno de mestrado) e Isabel Harb Manssour (professora orientadora), responsáveis pela pesquisa **Visualização do percurso acadêmico de estudantes ao longo do ensino superior**, estamos fazendo um convite para você participar como voluntário nesse estudo.

Esta pesquisa pretende como objetivo criar um modelo de visualização de dados que possibilite uma visão ampla do percurso acadêmico de um ou mais alunos ao longo dos semestres, indicando por meio de exploração de dados e análise estatística, indivíduos com tendência a não completarem seus cursos com sucesso. A ideia é não somente avaliar disciplinas com maior índice de reprovação, mas possibilitar uma visão contextualizada que inclui sua movimentação ao longo do tempo e entre cursos, além de dados estatísticos.

Acreditamos que esta pesquisa é importante porque a análise destes dados pode auxiliar os professores e administradores na tomada de decisões para minimizar as condições que levam os alunos à evasão.

Para sua realização será preciso fazer entrevistas seguindo um roteiro semiestruturado, pois mesmo nos baseando em estudos científicos sobre o tema, entendemos que a participação de profissionais que já realizam atividade analíticas visando a melhoria dos cursos de graduação e a diminuição da evasão pode enriquecer nosso trabalho, principalmente auxiliando no levantamento de requisitos. Por este motivo, realizamos o convite para sua participação neste trabalho.

Lembramos que será gravado uma sessão de áudio desta entrevista apenas para o pesquisador revisar se conseguiu captar todas as informações importantes durante a entrevista, posteriormente ao momento da conversa. Essa é uma ação que visa também reduzir o tempo de entrevista que seria utilizado para transcrever detalhes das respostas.

Sua participação constará de forma voluntária e você poderá pedir para retirar-se deste consentimento ou solicitar o encerramento da gravação a qualquer momento sem penalidades ou perda de despesas decorrentes de sua participação. Importante ressaltar que o objetivo deste estudo não é avaliar o participante, mas sim, entender os processos de trabalho, indicadores e ferramentas utilizadas em relação ao tema. O uso que faremos dos registros efetuados durante a

entrevista é estritamente limitado a atividades acadêmicas e será garantido o seu anonimato e confidencialidade.

Entendemos que há riscos mínimos durante essa atividade como: divulgação de dados confidenciais (quebra de sigilo) e desconforto, cansaço ou constrangimento durante a entrevista ou gravações de áudio. Você tem o direito de pedir uma indenização por qualquer dano que resulte da sua participação no estudo. Os benefícios que esperamos com o estudo são: Oferecer um modelo que propicie a visualização e a análise de dados de percurso de estudantes ao longo do tempo, permitindo auxiliar na identificação de estudantes com maior possibilidade de evasão.

Durante todo o período da pesquisa você tem o direito de esclarecer qualquer dúvida ou pedir qualquer outro esclarecimento, bastando para isso entrar em contato, com Fernando Lunardelli, no telefone celular (51) 99782-6726 ou Isabel H. Manssour no telefone celular (51) 99955-4948 a qualquer hora.

Em caso de algum problema relacionado com a pesquisa você terá direito à assistência gratuita que será prestada pelos pesquisadores e comitê de ética desta instituição.

Você tem garantido o seu direito de não aceitar participar ou de retirar sua permissão, a qualquer momento, sem nenhum tipo de prejuízo ou retaliação, pela sua decisão.

Se por algum motivo você tiver despesas decorrentes da sua participação neste estudo com transporte e/ou alimentação, você será reembolsado adequadamente pelos pesquisadores.

As informações desta pesquisa serão confidencias, e serão divulgadas apenas em eventos ou publicações científicas, não havendo identificação dos participantes, a não ser entre os responsáveis pelo estudo, sendo assegurado o sigilo sobre sua participação.

Ao assinar este termo de consentimento, você autoriza a gravação de áudio e não abre mão de nenhum direito legal que teria de outra forma.

Não assine este termo de consentimento a menos que tenha tido a oportunidade de fazer perguntas e tenha recebido respostas satisfatórias para todas as suas dúvidas.

A entrevista tem um tempo estimado de 60 até 90 minutos para sua conclusão.

Caso você tenha qualquer dúvida quanto aos seus direitos como participante de pesquisa, entre em contato com Comitê de Ética em Pesquisa da Pontifícia Universidade Católica do Rio Grande do Sul (CEP-PUCRS) em (51) 33203345, Av. Ipiranga, 6681/prédio 50 sala 703, CEP: 90619-900, Bairro Partenon, Porto Alegre – RS, e-mail: cep@pucrs.br, de segunda a sexta-feira das 8h às 12h e das 13h30 às 17h. O Comitê de Ética é um órgão independente constituído de

profissionais das diferentes áreas do conhecimento e membros da comunidade. Sua responsabilidade é garantir a proteção dos direitos, a segurança e o bem-estar dos participantes por meio da revisão e da aprovação do estudo, entre outras ações.

Se você concordar em participar deste estudo, você rubricará todas as páginas e assinará e datará duas vias originais deste termo de consentimento. Você receberá uma das vias para seus registros e a outra será arquivada pelo responsável pelo estudo.

Eu,_____, após a leitura deste documento e de ter tido a oportunidade de conversar com o pesquisador responsável, para esclarecer todas as minhas dúvidas, acredito estar suficientemente informado, ficando claro para mim que minha participação é voluntária e que posso retirar este consentimento a qualquer momento sem penalidades ou perda de qualquer benefício. Estou ciente também dos objetivos da pesquisa, dos procedimentos aos quais serei submetido, dos possíveis danos ou riscos deles provenientes e da garantia de confidencialidade e esclarecimentos sempre que desejar.

Diante do exposto expresso minha concordância de espontânea vontade em participar deste estudo, autorizando o uso, compartilhamento e publicação dos meus dados e informações de natureza pessoal para essa finalidade específica.

_____

Assinatura do participante da pesquisa ou de seu representante legal

_____

Assinatura de uma testemunha

DECLARAÇÃO DO PROFISSIONAL QUE OBTEVE O CONSENTIMENTO

Expliquei integralmente este estudo ao participante. Na minha opinião e na opinião do participante, houve acesso suficiente às informações, incluindo riscos e benefícios, para que uma decisão consciente seja tomada.

Data:_____

_____	_____
Fernando Lunardelli Aluno de Mestrado em Ciência da Computação PPGCC – Escola Politécnica	Isabel Harb Manssour Professora Orientadora PPGCC – Escola Politécnica