revista principia
Vol. 61, N. 3
REVISTA DE DIVULGAÇÃO CIENTÍFICA E TECNOLÓGICA DO INSTITUTO FEDERAL DE CIÊNCIA, EDUCAÇÃO E TECNOLOGIA DA PARAÍBA

DOI: http://dx.doi.org/10.18265/1517-0306a2022id7235
ARTIGO ORIGINAL

SUBMETIDO 07/09/2022
APROVADO 20/11/2022
PUBLICADO ON-LINE 03/12/2022
PUBLICADO 10/07/2024
EDITOR ASSOCIADO
Gilberto Reynoso Meza

Construção de um *data warehouse* para uma análise multidimensional da evasão escolar: estudo de caso da Plataforma Nilo Peçanha

Isleimar de Souza Oliveira [1]*
Fabio Gomes de Andrade [2]
Francisco Petrônio Alencar de Medeiros [3]
Camila Casimiro [4]
Fagne Rolim Farias [5]

[1] isleimar.oliveira@ifpb.edu.br
[2] fabio@ifpb.edu.br
[3] petronio@ifpb.edu.br
[4] camila.casimiro@academico.ifpb.edu.br
[5] fagne.farias@academico.ifpb.edu.br

Instituto Federal de Educação, Ciência e Tecnologia da Paraíba (IFPB), João Pessoa, Paraíba, Brasil

*Autor para correspondência.

RESUMO: Os índices elevados de evasão escolar representam um problema que tem atraído a atenção de muitos gestores e educadores ao redor do mundo. Esse problema acontece quando, por algum motivo, o aluno abandona o curso no qual está matriculado antes de sua conclusão. A evasão escolar gera uma série de prejuízos aos diversos atores envolvidos no processo educacional e pode ocorrer em diversos níveis e modalidades de ensino, por diversas causas. Uma forma de se combater a evasão escolar – e, consequentemente, os problemas causados por ela – consiste em identificar, de forma antecipada, alunos ou perfis de alunos com alto risco de evasão. A identificação prévia desses perfis pode auxiliar os gestores de políticas educacionais na tomada de decisão, como a elaboração de planos e a execução de ações que tentem evitar que os alunos abandonem os seus cursos. Este trabalho propõe a confecção de um banco de dados dimensional utilizado em um *data warehouse*, para análise das características dos alunos, relacionadas à evasão escolar. Os dados são oriundos da Plataforma Nilo Peçanha (PNP), adquiridos através de processos de extração, transformação e carga dos dados. Processo de Descoberta de Conhecimento em Bancos de Dados foi utilizado para analisar quais características estão relacionadas à evasão escolar de alunos de cursos de nível técnico e de graduação da rede federal de ensino de todo o Brasil, comparando os resultados com os dados do Instituto Federal de Educação, Ciência e Tecnologia da Paraíba (IFPB). As características relacionadas a faixa etária, renda per capita familiar e etnia apresentaram diferenças significativas quando comparados os alunos que abandonaram o curso e aqueles que o concluíram.

Palavras-chave: evasão escolar; *data warehouse*; PNP; processo de descoberta de conhecimento em bancos de dados.

Construction of a data warehouse for analysis of school dropout: Plataforma Nilo Peçanha case study

ABSTRACT: *School Dropout rates represent a problem that has attracted the attention of many managers and educators worldwide. This problem happens when, for some reason, the student leaves the course in which he is enrolled before its*

ISSN (on-line): 2447-9187

completion. School dropouts generate a series of damages to the various actors involved in the educational process, and these can occur at different levels and modalities of education for other causes. One way to combat school dropouts – and, consequently, the problems caused by it – is to identify in advance students or student profiles at a high risk of dropping out. The prior identification of these profiles can help managers of educational policies in decision-making, such as the elaboration of plans and the execution of actions that try to prevent students from abandoning their courses. In this work, we propose the creation of a dimensional database used in a data warehouse to analyze the characteristics of students related to school dropout. The data comes from the Nilo Peçanha Platform, acquired through data extraction, transformation, and loading processes. The Knowledge Discovery Process in the Database was used to analyze which characteristics are related to school dropout of students from technical and undergraduate courses of the federal education network throughout Brazil, comparing the results with data from the Instituto Federal de Educação, Ciência e Tecnologia da Paraíba (IFPB). The characteristics of age group, family per capita income, and ethnicity showed significant differences when comparing students who dropped out of the course with those who completed it.

Keywords: *data warehouse; knowledge discovery in databases; Nilo Peçanha Platform; school dropout.*

1 Introdução

Os altos índices de evasão escolar representam um problema que tem atraído a atenção de muitos gestores e educadores ao redor do mundo. Esse problema acontece quando, por algum motivo, o aluno abandona o curso no qual está matriculado antes de sua conclusão. A evasão pode ocorrer em diversos níveis, em diferentes modalidades de ensino e por diversas causas. Para Brito *et al*. (2015), os fatores que levam os alunos a deixarem os seus respectivos cursos estão relacionados, na maioria dos casos, a questões acadêmicas, dificuldades prévias em relação aos conteúdos e questões financeiras.

Especificamente no Brasil, a evasão escolar tem alcançado números preocupantes. Segundo dados fornecidos pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP, 2020), em sua Sinopse Estatística da Educação Superior, no ano de 2020 existia um total de 8.603.824 matrículas vinculadas aos cursos de graduação, nas modalidades presencial e a distância, no Brasil. Desse total de matrículas, excetuando-se os 1.229 alunos falecidos, um total de 3.744.522 alunos evadiram de seus cursos, o que representa aproximadamente 43,5% dos alunos matriculados.

A evasão escolar gera uma série de prejuízos aos diversos atores envolvidos no processo educacional. Para as instituições de ensino, sejam elas de natureza pública ou privada, a evasão pode gerar prejuízos financeiros. Nas instituições públicas, esse prejuízo ocorre porque alguns governos transferem parte dos recursos financeiros para as respectivas instituições de ensino de forma diretamente proporcional à quantidade de alunos matriculados. Assim, quando uma instituição tem um alto índice de evasão, ela pode perder uma quantidade significativa do seu orçamento. Nas instituições privadas, esse prejuízo ocorre porque a evasão afeta diretamente a receita proveniente das mensalidades pagas pelos alunos. Em relação aos alunos, a evasão escolar prejudica a sua qualificação profissional, o que, por sua vez, pode gerar dificuldades para que o aluno consiga oportunidades no mercado de trabalho ou limitar as suas possibilidades de crescimento profissional.

Além das questões já mencionadas, a evasão escolar gera problemas para os governos e para as comunidades nas quais os alunos estão inseridos. O problema causado aos governos é ocasionado pela ineficiência dos gastos públicos, pois, mesmo ao repassar parte dos recursos às instituições de ensino de forma proporcional ao número de alunos, os governos ainda precisam arcar com gastos fixos como o salário dos professores e demais servidores da instituição e, do mesmo modo, com a manutenção de toda a infraestrutura utilizada para as atividades de ensino. Ademais, altos índices de evasão dificultam a qualificação da mão de obra, o que, por sua vez, pode gerar problemas de desemprego e tornar as regiões nas quais os alunos estão inseridos pouco atraentes para o investimento em novos empreendimentos que exijam essa qualificação (Prim; Fávero, 2013).

Uma forma de se combater a evasão escolar e, consequentemente, os problemas causados por ela consiste em identificar, de forma antecipada, alunos ou perfis de alunos com alto risco de evasão. A identificação prévia desses perfis pode auxiliar gestores educacionais em processos de tomada de decisão, assim como é possível elencar uma série de ações e de políticas para melhorar o acompanhamento desses alunos, de forma a evitar que eles abandonem os seus cursos (Romero *et al*., 2010).

Para Dore e Lüscher (2011), a evasão escolar é um fenômeno afetado por múltiplos fatores, necessitando análises em diferentes aspectos, sejam individuais, familiares, institucionais, sociais e inclusive comunitários, sendo o maior problema associado a escassez de informações sobre o assunto. Conforme apontado por Veiga e Bergiante (2016), existe uma grande variedade de fatores responsáveis pela permanência do aluno na escola, e a identificação desses fatores está ligada diretamente à análise do conjunto de alunos.

Nos últimos anos, vários trabalhos têm sido desenvolvidos na área da ciência da computação para reconhecer os fatores que contribuem para a evasão escolar (Kabashi; Shabani; Caka, 2022; Singh; Alhulail, 2022). Muitos desses trabalhos utilizam técnicas de mineração de dados para detectar alunos com alto risco de evasão (Kaur; Singh; Josan, 2015; Mnyawami; Maziku; Mushi, 2022; Moliner; Alegre; Lorenzo-Valentin, 2022; Xu *et al*., 2022). A mineração de dados tem como objetivo adquirir informação a partir da análise de um grande conjunto de dados, mediante técnicas e algoritmos que buscam por padrões e tendências presentes nesses dados (Castro; Ferrari, 2016).

Silva e Mendonça (2021) realizaram uma análise da taxa de evasão de cursos de licenciatura das instituições federais de ensino presentes na Plataforma Nilo Peçanha no cenário da pandemia causada pelo SARS-CoV-2 (covid-19). Para a análise exploratória, foi utilizado o fluxo de extração, tratamento e processamento dos dados, sendo essa uma etapa para o processo de Descoberta de Conhecimento em Bancos de Dados (no inglês *Knowledge Discovery in Databases*, KDD), com o auxílio dos softwares Jamovi, Excel, Enterprise Resource Planning CONECTA TOTVS e Business Intelligence Power BI. Quanto à análise sobre a retenção e evasão escolar, o trabalho identificou que existe um equilíbrio em relação aos gêneros e que, apesar do número significativo de evasão dos cursos superiores, não houve um aumento da retenção ou da evasão durante o período de pandemia de covid-19.

Santos, Simon e Pinto (2020) realizaram uma análise quantitativa sobre o fenômeno da retenção e evasão escolar nos cursos superiores do Campus Júlio de Castilho do Instituto Federal Farroupilha, utilizando dados da Coordenação de Registro Acadêmico da instituição em conjunto com os dados disponibilizados pela Plataforma Nilo Peçanha. Para a organização e análise exploratória dos dados foi utilizado o software de planilha eletrônica Calc do LibreOffice. Como resultado foi observado o número significativo de alunos evadidos frente ao total de alunos matriculados, tendo uma taxa de retenção do ciclo de 23,43%, uma taxa de evasão do ciclo de 41,71% e índice de eficiência acadêmica de 45,5%.

Para os trabalhos analisados, não é possível agrupar os dados em múltiplas dimensões (utilizando-se diferentes granularidades), sendo difícil replicar os resultados em múltiplos conjuntos de dados. Ademais, nos trabalhos não é possível analisar os atributos para a evasão escolar em diferentes dimensões inerentes aos dados, tais como a faixa etária, gênero/sexo, condições socioeconômicas, entre outras. Para resolver essas limitações, este artigo apresenta um framework no qual a modelagem dimensional e os conceitos de inteligência de negócios foram usados para realizar uma análise mais profunda sobre o conjunto de dados abertos disponibilizados.

O objetivo deste trabalho é investigar e desenvolver um modelo dimensional para a análise de dados de evasão em todo o Brasil, permitindo que eles sejam visualizados a partir de diversos atributos. Como um estudo de caso da aplicação do modelo proposto, foram utilizados dados sobre os cursos de nível técnico e de graduação da rede federal de ensino de todo o Brasil. Os resultados em nível nacional foram comparados com os dados do Instituto Federal de Educação, Ciência e Tecnologia da Paraíba (IFPB) disponíveis no portal de dados abertos da Plataforma Nilo Peçanha (PNP), o qual publica dados sobre as instituições de ensino que compõem a rede.

O restante deste artigo está organizado da seguinte forma: a seção 2 descreve o referencial teórico, apresentando as tecnologias utilizadas na elaboração do trabalho; na seção 3 é descrito o método da pesquisa e os processos para aquisição dos dados e construção do *data warehouse*; na seção 4 são exibidos resultados da análise a partir das várias dimensões dos dados; a seção 5 conclui o artigo e apresenta as considerações finais.

2 Referencial teórico

Os dados analisados neste trabalho foram agrupados utilizando um *data warehouse* (DW), que significa armazém de dados. O DW surgiu com a necessidade de se agrupar dados corporativos que estão espalhados em diferentes máquinas e sistemas operacionais, sendo o DW um banco de dados especializado, que utiliza conceitos totalmente diferentes, a partir dos quais são construídos os sistemas de ambiente operacional ou OLTP (*Online Transaction Processing* ou Processamento de Transações em Tempo Real). Essa tecnologia é muito utilizada em empresas que desejam alcançar vantagem competitiva, auxiliando gestores no processo de tomada de decisão, os quais têm acesso a dados que representam fatos ocorridos, e não apenas a especulações ou previsões, reduzindo, assim, a probabilidade de erro (Kemczinski *et al*., 2003). Inmon (2002) define um DW como uma coleção de dados orientados por assuntos, integrados, variáveis com o tempo e não voláteis, para dar suporte ao processo de tomada de decisão.

Uma das técnicas mais utilizadas para construção do projeto lógico de dados de um DW é a modelagem dimensional. De acordo com Kimball e Ross (2013), a modelagem dimensional é uma técnica de projeto de banco de dados definida para dar suporte às consultas feitas pelos usuários finais de um DW. Sua principal característica é o alto desempenho na execução de consultas, e, para isso, esse tipo de modelagem não busca atender à normalização dos dados, em especial, à terceira forma normal. A modelagem dimensional tem por intuito criar uma estrutura simples que seja entendida por todos os envolvidos no negócio e que seja resiliente a mudanças. Os três principais elementos que formam esse tipo de modelagem são: fatos, dimensões e métricas.

Os dados persistidos em um DW são oriundos de diversas fontes, porém os dados não podem ser transferidos diretamente da fonte para o DW, sendo necessário realizar um processo de Extração, Transformação e Carga (no inglês *Extract, Transform and Load* – ETL). O ETL é um processo dividido em três etapas; a primeira etapa é a extração

dos dados, que pode ocorrer a partir de diversas fontes. A segunda é a transformação, na qual os dados são modificados, aplicando técnicas de higienização, padronização, filtragem e criação do dicionário. Por fim, na etapa de carga os dados são persistidos em uma base de dados (Ferreira *et al*., 2010; Sakka *et al*., 2021).

Os dados persistidos na base de dados do DW são amplamente utilizados em processos de Descoberta de Conhecimento em Bancos de Dados (no inglês *Knowledge Discovery in Databases* – KDD), que têm como propósito extrair informação em grandes conjuntos de dados buscando padrões explicáveis neles, permitindo a interpretação e extrapolação para eventos futuros (Fayyad; Piatetsky-Shapiro; Smyth, 1996). Para Baker, Isotani e Carvalho (2011), o processo de KDD extrai conhecimento implícito a partir de um conjunto de dados volumosos, possibilitando obter *insights* que auxiliem em tomadas de decisões. Segundo Fayyad, Piatetsky-Shapiro e Smyth (1996), o KDD está dividido nas etapas de seleção de dados, pré-processamento, transformação, mineração de dados e interpretação dos resultados.

A principal etapa do KDD é a mineração de dados (em inglês *Data Mining* – DM), que aplica análise a grandes quantidades de dados para obter conhecimento, agregar significado e utilidade aos dados (Castro; Ferrari, 2016). A DM é denominada Mineração de Dados Educacional (em inglês *Educational Data Mining* – EDM) quando os dados são oriundos do contexto educacional. A EDM torna-se cada vez mais reconhecida como uma ferramenta emergente que se concentra no desenvolvimento de métodos para explorar os dados provenientes do cenário educacional (Romero *et al*., 2010).

3 Método da pesquisa

Nesta seção são descritos a origem dos dados utilizados, a sistemática para construção do DW, descrição do modelo dimensional e etapas de processamento dos dados, que abrangem a extração, tratamento e persistências, e por fim as ferramentas utilizadas para análise dos dados.

3.1 Obtenção dos dados para a pesquisa

Para execução do projeto foi utilizada a base de dados aberta da Plataforma Nilo Peçanha (PNP) do governo federal, na qual são disponibilizados anualmente, desde 2018, microdados referentes às instituições de ensino que compõem a rede federal de ensino, tais como os institutos e as universidades federais. Entre os dados disponibilizados pela plataforma, foram utilizados nesta pesquisa os "*Microdados Matrícula*", que possuem informações relacionadas às matrículas de alunos, cursos e instituições. Cada linha desse arquivo contém o registro da situação acadêmica de um aluno da rede federal. Para cada aluno é disponibilizada uma série de informações, como: sexo, etnia, faixa de renda, idade, tipo do curso, eixo tecnológico, modalidade de ensino, vagas ofertadas, turno, instituição de ensino, município, unidade federativa e situação do aluno no curso.

Os microdados disponibilizados em cada ano se referem à situação dos alunos no ano anterior – por exemplo, os microdados de 2018 são referentes à situação dos alunos em 2017, enquanto os correspondentes ao ano de 2018 são representados no conjunto de dados disponibilizado em 2019 e assim sucessivamente. Durante a execução deste trabalho, foram usados os microdados da PNP disponibilizados nos anos de 2019 e 2020, pois os dados de 2018 apresentam uma quantidade reduzida de campos em referência

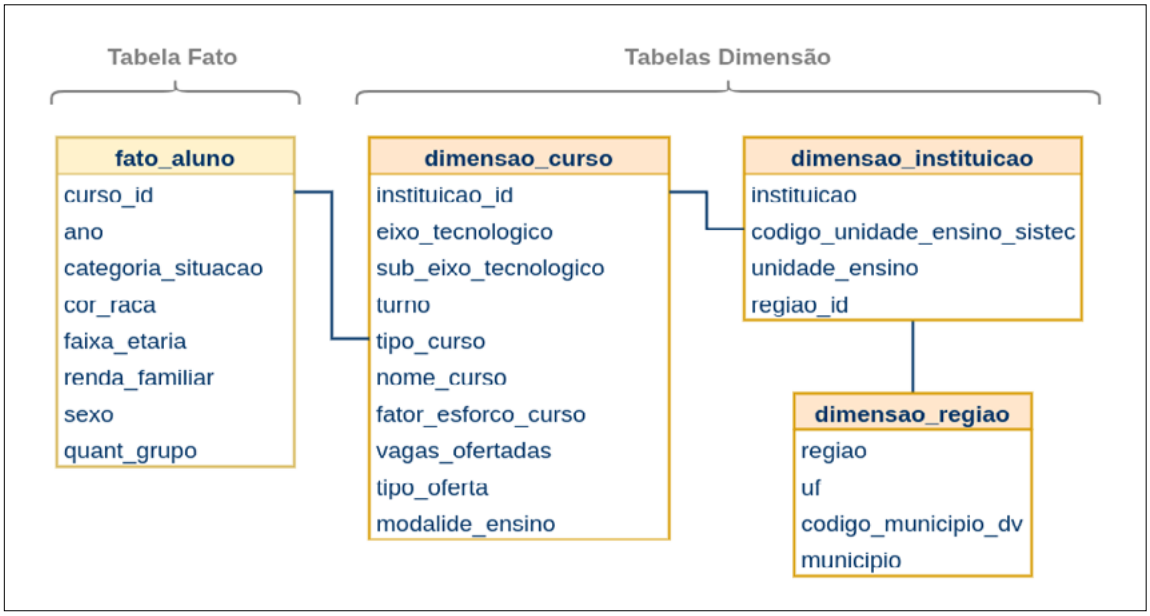
aos demais anos, de modo que sua utilização produziria um banco de dados com campos vazios em muitos registros.

3.2 Definição do esquema dimensional

Uma vez definidos os dados a serem usados na pesquisa, foi criado um banco de dados local para o armazenamento. A criação desse banco de dados foi importante para agilizar a realização das pesquisas e, por consequência, o processo de análise da evasão escolar.

Antes da implementação, foi criado um esquema lógico para definir a estrutura a ser utilizada para a organização dos dados. O esquema lógico proposto é baseado no modelo dimensional. A opção pela utilização de um esquema desse tipo foi motivada com base na intenção de organizar os dados como um DW, de forma a permitir que os dados de evasão pudessem ser analisados a partir de várias dimensões e em diversos níveis de agrupamento. O esquema lógico proposto para a implementação do trabalho é composto por uma tabela fato, chamada *fato_aluno*, e três tabelas de dimensão, sendo elas *dimensao_curso*, *dimensao_instituicao* e *dimensao_regiao*, conforme mostrado na Figura 1.

Figura 1 ▼
Diagrama do esquema dimensional do *data warehouse*.
Fonte: dados da pesquisa



Esse esquema lógico foi implementado em um banco de dados NoSQL Apache Cassandra[1] versão 4.0.4. O Cassandra foi escolhido por ser, conforme descrito por Carpenter e Hewitt (2020), um banco de dados *open source*, distribuído e escalável, altamente disponível, tolerante a falhas e orientado a linhas, adequado para cargas de trabalho transacionais volumosas e amplamente utilizado em DWs.

[1] Apache Cassandra Open Source NoSQL Database. Disponível em: https://cassandra.apache.org/.

A tabela *fato_aluno* armazena as informações sobre os diferentes perfis de alunos que compõem a rede federal de ensino. Para cada perfil de aluno, são registrados os seguintes atributos: *curso_id*, que identifica o curso no qual o perfil de aluno realizou a matrícula; *ano*, que representa o ano de referência do registro da informação do aluno; *categoria_situacao*, que agrupa o *status* da situação acadêmica do aluno, podendo assumir os valores Em curso, Concluinte e Evadido. Outros atributos são: *cor_raca*, que registra o grupo étnico declarado pelo aluno na matrícula; *faixa_etaria*, que representa o grupo etário no qual o aluno está incluído no momento do registro, podendo ser: menor de 15 anos,

15 a 19 anos, 20 a 24 anos, 25 a 29 anos, 30 a 34 anos, 35 a 39 anos, 40 a 44 anos, 45 a 49 anos, 50 a 54 anos, 55 a 59 anos e maior de 60 anos; *renda_familiar*, que registra o valor da Renda Familiar por Pessoa (RFP) com relação ao valor do salário mínimo declarado pelo aluno, cujos grupos já são definidos no arquivo obtido na PNP, sendo composta pelos seguintes valores: "0 < RFP <= 0,5", "0,5 < RFP <= 1", "1 < RFP <= 1,5", "1,5 < RFP <= 2,5", "2,5 < RFP <= 3,5", "RFP > 3,5" e "Não declarada". Os atributos seguintes são: *sexo*, que registra o gênero declarado pelo aluno; e *quant_grupo*, que possui a quantidade de registros agrupados conforme as características descritas anteriormente.

A tabela *dimensao_curso* é responsável por qualificar os cursos aos quais os alunos estão relacionados. Para cada curso são registradas informações como o nome do curso, o eixo tecnológico, o subeixo tecnológico, o turno, o tipo de curso, o fator de esforço, a quantidade de vagas ofertadas, o tipo de oferta e a modalidade de ensino. Além disso, cada curso tem um atributo chamado *instituicao_id*, relacionado à instituição. A hierarquia definida nessa tabela permite que os dados sejam analisados em diferentes níveis de agrupamento. Por exemplo, é possível analisar os dados de evasão para toda a rede federal (o maior nível de agrupamento possível), apenas para os cursos de tecnologia, apenas para os cursos de um determinado eixo e subeixo tecnológico ou apenas para um curso específico, que representa o menor nível de agrupamento dessa dimensão.

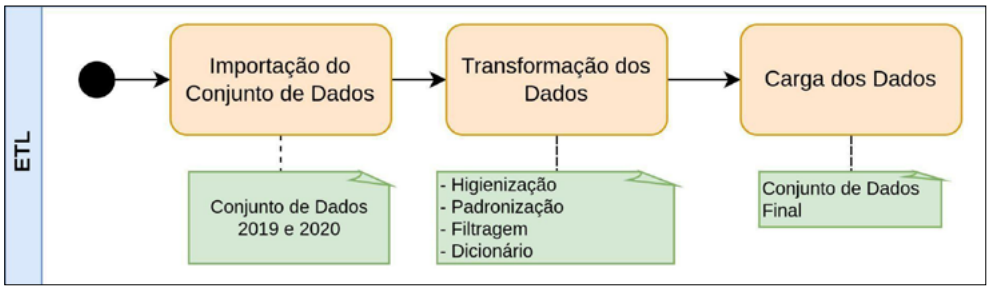
A tabela *dimensao_instituicao* representa a dimensão que abstrai as informações acerca das instituições de ensino. Para cada tupla dessa tabela são armazenados o código da instituição, a instituição (por exemplo, Instituto Federal da Paraíba) e a unidade de ensino (por exemplo, o Campus Cajazeiras). A hierarquia definida nessa dimensão permite que sejam realizadas análises em diferentes níveis de granularidade. Por exemplo, pode ser feita uma análise dos dados de evasão para toda uma instituição de ensino ou apenas para um *campus* específico.

Finalmente, *dimensao_regiao* é uma tabela de dimensão criada para permitir a análise dos dados com base no perfil geográfico das instituições de ensino. Para cada tupla dessa tabela são armazenados o código do município, o nome do município, o nome da unidade federativa e o nome da região. A hierarquia definida nessa dimensão permite que sejam realizadas análises em nível de município, estado e região.

3.3 Ferramenta ETL

Uma vez definidos os dados para a implementação e criação do banco de dados, a próxima etapa consistiu no desenvolvimento da ferramenta ETL, que é responsável por ler os dados obtidos na plataforma PNP, transformá-los para o esquema lógico proposto e carregá-los no banco de dados. Esse processo é realizado em três etapas, conforme descrito na Figura 2.

Figura 2 ▶
Diagrama de atividades do processo de ETL.
Fonte: dados da pesquisa



A primeira etapa do processo de ETL corresponde à importação do conjunto de dados. Nela, os dados da PNP foram extraídos, descompactados e armazenados localmente com a utilização de um arquivo de *script* escrito na linguagem de programação Python[2] (versão 3.8.10). Todos os dados utilizados neste trabalho foram disponibilizados pela plataforma no formato .csv. Os dados referentes ao ano de 2018 apresentaram limitações em relação à quantidade de informações disponíveis, uma vez que o arquivo contendo esses dados não possui alguns campos presentes nos demais períodos. Assim, com o objetivo de dispor de uma base maior tanto na quantidade de registros quanto na quantidade de atributos, decidiu-se por não incluir os dados do período 2018, integrando no banco de dados apenas os dados dos anos de 2019 e 2020.

[2] Welcome to Python.org. Disponível em: https://www.python.org.

Uma vez importados, os dados são transformados para o esquema lógico definido para o seu armazenamento. Devido ao grande número de registros existentes em cada arquivo e ao esforço para a análise da integridade dos dados, a transformação é a etapa mais dispendiosa do processo de ETL, pois os dados necessitam passar por diversos processos, como higienização, padronização, filtragem e geração de dicionário.

Na etapa de higienização, durante a análise exploratória dos dados, foram identificadas inconsistências, como a existência de valores nulos, de maneira que esses registros foram removidos da base de dados final. No campo "idade", foi verificada a existência de registros como valores superiores a 12.000 anos ou idades negativas; nesses casos, os registros foram excluídos, aplicando filtros para remover os valores fora do intervalo de 10 a 90 anos.

Na etapa de padronização, os registros que referenciam situações iguais, porém com formatações diferentes, são convertidos para um mesmo padrão, pois alguns campos apresentam, nas duas bases de dados, formatos diferentes mesmo quando se referem aos mesmos valores. Por exemplo, o campo sexo, enquanto na base de 2019 apresenta os valores "F" e "M", na base de 2020 apresenta como valores "Masculino" e "Feminino". Desse modo, para realizar a integração das duas bases de dados, fez-se necessário transformar os valores desses campos para um mesmo formato visando à padronização da base de dados final.

O campo "ano" foi incluído, derivando sua origem do período da base de dados da qual foi extraído. Na etapa de criação dos dicionários, os campos de valores relacionados que apresentavam nomes diferentes foram modificados para que todas as instâncias pudessem ser acessadas utilizando o mesmo atributo, de forma a padronizar os nomes dos atributos para serem utilizados na integração das duas bases de dados selecionadas.

Na etapa de filtragem foram aplicados filtros de acordo com as regras de negócio, sendo o campo *categoria_situacao* filtrado com o propósito de selecionar apenas os registros que apresentavam os valores "Concluinte" ou "Evadido", já que os registros "Em Curso" não possuem um dos estados desejados para o objetivo do trabalho, que é identificar se o aluno concluiu ou evadiu. Dessa forma não foram incluídos na análise os registros cujo campo *categoria_situação* apresentavam o valor "Em Curso", para fins de identificação da porcentagem de conclusão em cada grupo. Por último, o filtro do campo *tipo_curso* restringe os valores a "Bacharelado", "Tecnologia" e "Licenciatura", quando delimita o foco para os dados de cursos superiores, e a "Técnicos", quando para os cursos técnicos.

4 Resultados da análise do *data warehouse*

Uma vez construído o DW, os dados sobre evasão foram analisados a partir de várias dimensões. Quanto à instituição de ensino, foram selecionados os alunos das instituições públicas no Brasil e alunos oriundos do IFPB. Quanto aos tipos de curso,

foram selecionados dois grupos, referentes aos alunos de curso técnico e de curso superior. Quanto às características dos alunos, foram analisados os atributos de cor/raça, faixa etária, renda e sexo como forma de exemplificar a utilização do DW para dados.

4.1 Análise por tipo do curso

Primeiramente, foi realizada uma análise em âmbito nacional com base no tipo de curso. Os dados obtidos nessa análise são mostrados na Tabela 1. A primeira coluna dessa tabela indica o tipo de curso. As demais colunas indicam, respectivamente, o total de alunos que concluíram ou evadiram do curso, o total de alunos concluintes e o total de alunos evadidos. É importante ressaltar que os dados sobre educação infantil foram omitidos da Tabela 1, uma vez que a quantidade de registros sobre esse tipo de curso é muito pequena.

Tabela 1 ▼
Quantitativo de registros nos cursos nacionais e porcentagens de conclusão e evasão.
Fonte: dados da pesquisa

Tipo de curso	Total	Concluintes	Evadidos
Técnico	307.378 (42,41%)	158.976 (51,72%)	148.402 (48,28%)
Formação Inicial e Continuada	273.804 (37,78%)	173.108 (63,22%)	100.696 (36,78%)
Graduação	110.717 (15,28%)	76.527 (69,12%)	34.190 (30,88%)
Especialização	22.690 (3,13%)	12.563 (55,37%)	10.127 (44,63%)
Ensino Médio	3.616 (0,50%)	2.120 (58,63%)	1.496 (41,37%)
Ensino Fundamental	3.296 (0,45%)	2.873 (87,17%)	423 (12,83%)
Mestrado	3.129 (0,43%)	2.364 (75,55%)	765 (24,45%)
Doutorado	108 (0,01%)	89 (82,41%)	19 (17,59%)
Total	**724.738**	**428.620 (59,14%)**	**296.118 (40,86%)**

Ao se analisar a Tabela 1, percebe-se que os maiores percentuais de conclusão ocorreram no ensino fundamental (87,17%), doutorado (82,41%) e mestrado (75,55%). Por outro lado, os maiores índices de evasão foram observados nos cursos técnicos (48,28%), especialização (44,63%) e ensino médio (41,37%). Todos os cursos possuem porcentagem de conclusão superior a 50%; desse modo, mais da metade dos alunos concluem os cursos.

A Tabela 2 indica o quantitativo de registros por tipo de curso com base apenas em alunos do IFPB. Nela, percebe-se que a quantidade total de alunos que finalizaram os seus cursos (concluindo ou evadindo) no IFPB foi de 14.333, sendo 8.633 evadidos e 5.700 concluintes.

Tabela 2 ▼
Quantitativo de registros nos cursos do IFPB e porcentagens de conclusão e evasão.
Fonte: dados da pesquisa

Tipo de curso	Quantidade	Concluintes	Evadidos
Técnico	7.638 (53,29%)	4.117 (53,90%)	3.521 (46,10%)
Graduação	4.514 (31,49%)	3.138 (69,52%)	1.376 (30,48%)
Formação Inicial e Continuada	1.881 (13,12%)	1.164 (61,88%)	717 (38,12%)
Especialização	273 (1,90%)	193 (70,70%)	80 (29,30%)
Mestrado	27 (0,19%)	21 (77,78%)	6 (22,22%)
Total	**14.333**	**8.633 (60,23%)**	**5.700 (39,77%)**

Ao analisar a Tabela 2, é possível observar que o IFPB possui uma porcentagem de conclusão geral de 60,23%, o que representa um aumento de 1,09 ponto percentual em comparação ao percentual nacional, que é de 59,14%. Apresentam também um percentual de conclusão maior que a média nacional, em termos percentuais, os cursos técnicos (+2,18%), cursos de graduação (com +0,40%), mestrado (+9,39%) e especialização (+15,33%). Já os cursos de Formação Inicial e Continuada do IFPB têm um percentual de conclusão 1,34% menor do que a média nacional.

As porcentagens em relação à quantidade de registro dos alunos nos cursos no IFPB são 53,29% nos cursos técnicos, 31,49% nos cursos superiores, 13,12% nos cursos FIC, 1,90% na especialização e 0,19% no mestrado. Devido à pequena porcentagem de registros dos cursos FIC, especialização e mestrado, foram analisados nas etapas seguintes apenas os dados dos cursos técnicos e superiores.

4.2 Análise por cor/raça

A Tabela 3 mostra os resultados decorrentes da análise dos dados sobre a cor/raça no DW, para o conjunto de dados dos cursos técnicos no Brasil. É possível constatar que os alunos das raças "Amarela", "Branca", "Não Declarada" e "Parda" apresentaram, respectivamente, um percentual de conclusão superior a 50%. Por outro lado, os resultados mais baixos foram apresentados pelos grupos "Preta" e "Indígena", mas, ainda assim, com resultados superiores a 40%.

Tabela 3 ▶
Quantitativo de registros na dimensão cor/raça nos cursos técnicos no Brasil.
Fonte: dados da pesquisa

Dimensão cor/raça	Quantidade	Concluintes	Evadidos
Parda	108.806 (35,40%)	54.721 (50,29%)	54.085 (49,71%)
Não declarada	99.261 (32,29%)	50.979 (51,36%)	48.282 (48,64%)
Branca	70.089 (22,80%)	39.182 (55,90%)	30.907 (44,10%)
Preta	23.036 (7,49%)	10.691 (46,41%)	12.345 (53,59%)
Amarela	4.313 (1,40%)	2.580 (59,82%)	1.733 (40,18%)
Indígena	1.873 (0,61%)	823 (43,94%)	1.050 (56,06%)
Total	**307.378**	**158.976 (51,72%)**	**148.402 (48,28%)**

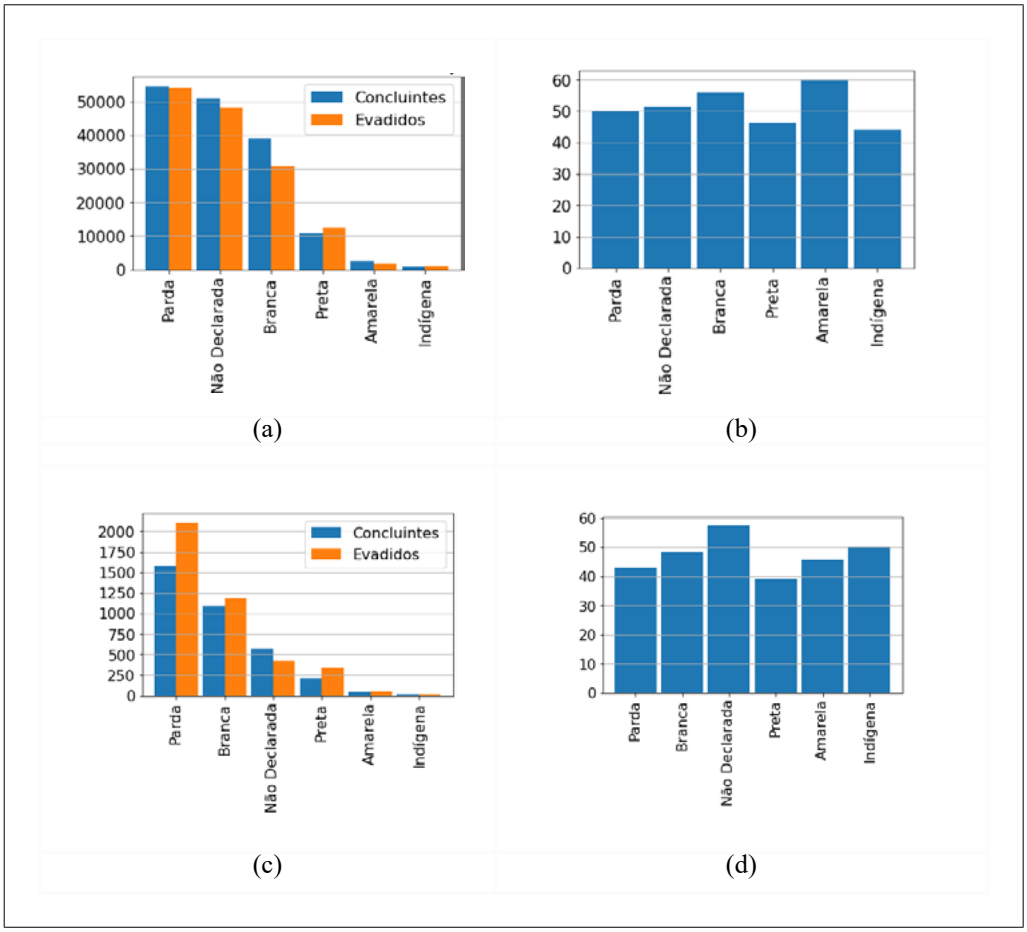
A Tabela 4 apresenta o quantitativo de registros dos cursos técnicos no IFPB, em que apenas os alunos "Não Declarados" obtiveram porcentagem de conclusão superior a 50,00%. Já os indígenas apresentaram um percentual de conclusão de exatamente 50,00%, enquanto os alunos do grupo "Preta" apresentaram uma porcentagem de conclusão inferior a 40,00%, um valor inferior à média nacional.

Tabela 4 ▶
Quantitativo de registros na dimensão cor/raça nos cursos técnicos no IFPB.
Fonte: dados da pesquisa

Dimensão cor/raça	Quantidade	Concluintes	Evadidos
Parda	3.689 (48,30%)	1.579 (42,80%)	2.110 (57,20%)
Branca	2.274 (29,77%)	1.094 (48,11%)	1.180 (51,89%)
Não declarada	999 (13,08%)	574 (57,46%)	425 (42,54%)
Preta	550 (7,20%)	215 (39,09%)	335 (60,91%)
Amarela	96 (1,26%)	44 (45,83%)	52 (54,17%)
Indígena	30 (0,39%)	15 (50,00%)	15 (50,00%)
Total	**7.638**	**3.521 (46,10%)**	**4.117 (53,90%)**

A Figura 3 apresenta os gráficos construídos a partir da análise das quantidades de registros em função da porcentagem de conclusão dos alunos dos cursos técnicos no Brasil e no IFPB. Nela, é possível observar inversão entre a proporção de estudantes cuja Cor/Raça foi registrada como "Não Declarada" e os que declararam "Branca" quando comparados os conjuntos de dados nacionais e do IFPB.

Figura 3 ▶
Análise dos cursos técnicos em relação à cor/raça. (a) Quantidade por categoria no Brasil. (b) Porcentagem de concluintes no Brasil. (c) Quantidade por categoria no IFPB. (d) Porcentagem de concluintes no IFPB.
Fonte: dados da pesquisa



A Tabela 5 (próxima página) aponta os resultados decorrentes da análise quantitativa dos dados sobre a cor/raça para o conjunto de dados dos cursos superiores no Brasil. Observa-se que nenhum grupo apresenta um percentual de conclusão superior a 34%. Os grupos com percentuais de conclusão mais baixos correspondem às raças "Parda" e "Preta", com resultados inferiores a 30%.

Tabela 5 ▶
Quantitativo de registros na dimensão cor/raça nos cursos superiores no Brasil.
Fonte: dados da pesquisa

Dimensão cor/raça	Quantidade	Concluintes	Evadidos
Parda	35.953 (32,47%)	10.344 (28,77%)	25.609 (71,23%)
Branca	33.633 (30,38%)	11.077 (32,93%)	22.556 (67,07%)
Não declarada	31.565 (28,51%)	10.105 (32,01%)	21.460 (67,99%)
Preta	7.975 (7,20%)	2.152 (26,98%)	5.823 (73,02%)
Amarela	1.174 (1,06%)	373 (31,77%)	801 (68,23%)
Indígena	417 (0,38%)	139 (33,33%)	278 (66,67%)
Total	**110.717**	**34.190 (30,88%)**	**76.527 (69,12%)**

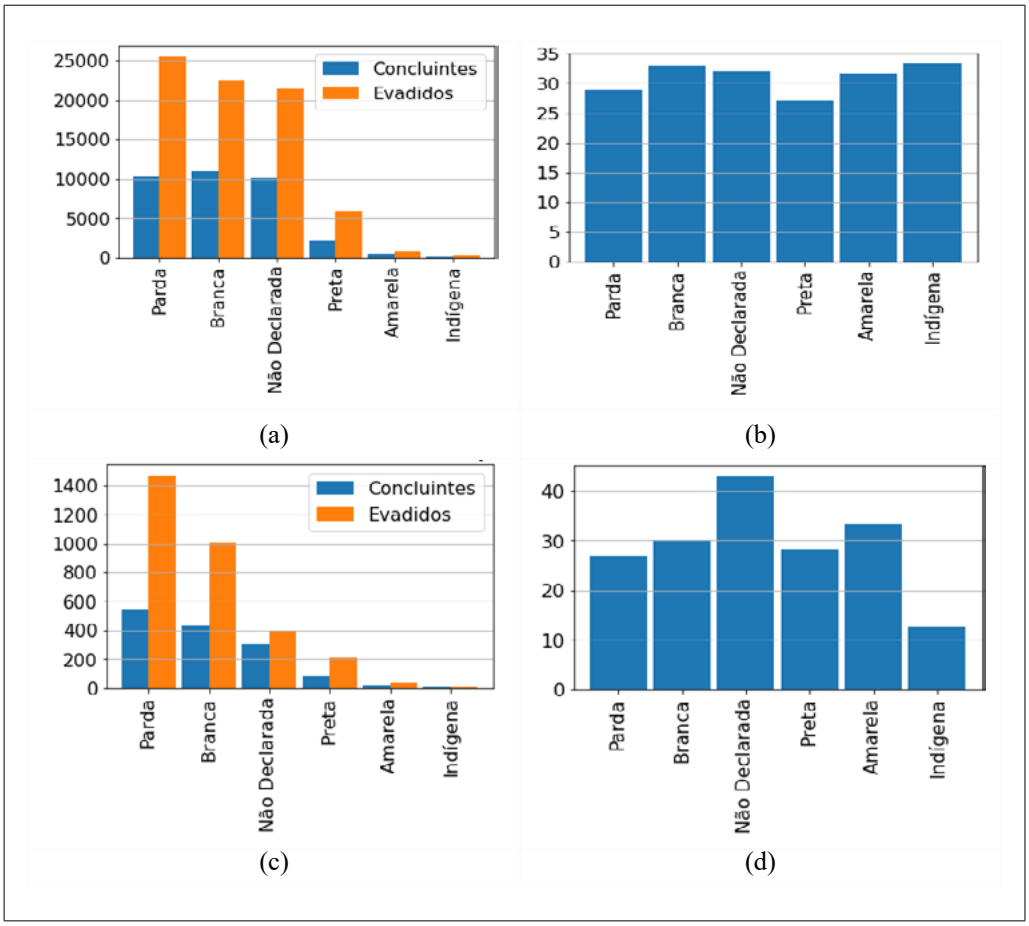
A Tabela 6 indica o quantitativo de registros dos cursos superiores no IFPB. Nos resultados é possível observar que, com exceção dos alunos que não declararam a cor/raça, todos os demais grupos obtiveram percentual de conclusão inferior a 40,00%. Os resultados mais baixos são apresentados pelo grupo "Indígenas", com porcentagem inferior a 13%, um valor bem abaixo da média nacional, que é de 33,33%.

Tabela 6 ▶
Quantitativo de registros na dimensão cor/raça nos cursos superiores no IFPB.
Fonte: dados da pesquisa

Dimensão cor/raça	Quantidade	Concluintes	Evadidos
Parda	2.011 (44,55%)	538 (26,75%)	1.473 (73,25%)
Branca	1.438 (31,86%)	432 (30,04%)	1.006 (69,96%)
Não Declarada	703 (15,57%)	302 (42,96%)	401 (57,04%)
Preta	297 (6,58%)	84 (28,28%)	213 (71,72%)
Amarela	57 (1,26%)	19 (33,33%)	38 (66,67%)
Indígena	8 (0,18%)	1 (12,50%)	7 (87,50%)
Total	**4.514**	**1.376 (30,48%)**	**3.138 (69,52%)**

A Figura 4 apresenta os gráficos da análise para os cursos superiores no Brasil e no IFPB. Quando comparados os dados nacionais e os do IFPB, é possível constatar uma diferença considerável entre os percentuais de conclusão nos dois conjuntos de dados. O grupo "Indígena" apresenta as maiores diferenças entre os dados nacionais e os do IFPB, podendo ser causadas pela baixa quantidade de registros existentes.

Figura 4 ▶
Análise dos cursos superiores em relação à cor/raça.
(a) Quantidade por categoria no Brasil.
(b) Porcentagem de concluintes no Brasil.
(c) Quantidade por categoria no IFPB.
(d) Porcentagem de concluintes no IFPB.
Fonte: dados da pesquisa



4.3 Análise por faixa etária

Os resultados decorrentes da análise quantitativa dos dados sobre faixa etária no DW para o conjunto de dados dos cursos superiores no Brasil são exibidos na Tabela 7. É possível constatar que o melhor resultado é apresentado pelo grupo de "25 a 29 anos", com percentual de conclusão superior a 40%. Por outro lado, os grupos "Menor de 15 anos" e "15 a 19 anos" tiveram percentuais inferiores a 13%. Os demais grupos apresentaram percentuais próximos de 30%.

Tabela 7 ▶
Quantitativo de registros na dimensão faixa etária nos cursos superiores no Brasil.
Fonte: dados da pesquisa

Dimensão faixa etária	Quantidade	Concluintes	Evadidos
Menor de 15 anos	8 (0,01%)	1 (12,50%)	7 (87,50%)
15 a 19 anos	9.584 (8,66%)	65 (0,68%)	9.519 (99,32%)
20 a 24 anos	40.285 (36,39%)	13.008 (32,29%)	27.277 (67,71%)
25 a 29 anos	27.180 (24,55%)	10.964 (40,34%)	16.216 (59,66%)
30 a 34 anos	14.396 (13,00%)	4.429 (30,77%)	9.967 (69,23%)
35 a 39 anos	8.691 (7,85%)	2.468 (28,40%)	6.223 (71,60%)
40 a 44 anos	4.870 (4,40%)	1.422 (29,20%)	3.448 (70,80%)
45 a 49 anos	2.761 (2,49%)	859 (31,11%)	1.902 (68,89%)
50 a 54 anos	1.631 (1,47%)	520 (31,88%)	1.111 (68,12%)
55 a 59 anos	863 (0,78%)	300 (34,76%)	563 (65,24%)
Maior de 60 anos	448 (0,40%)	154 (34,38%)	294 (65,63%)
Total	**110.717**	**34.190 (30,88%)**	**76.527 (69,12%)**

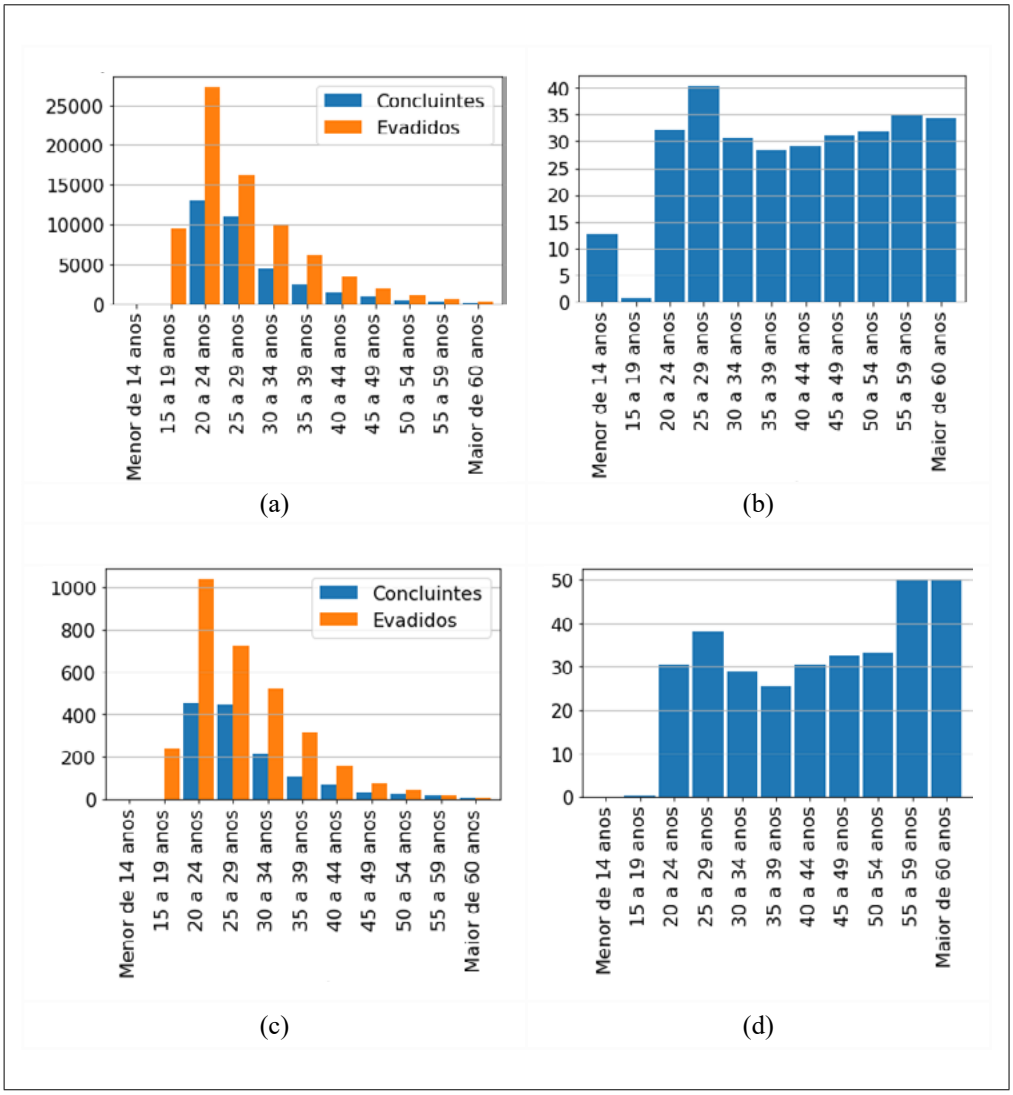
A Tabela 8 apresenta o quantitativo de registros dos cursos superiores no IFPB. Nos resultados verifica-se que, com exceção dos grupos "55 a 59 anos" e "Maior de 60 anos", que apresentam porcentagem de conclusão igual a 50%, os demais grupos mostram um percentual de conclusão inferior a 40%. Também é possível constatar que apenas os alunos dos grupos "Menor de 15 anos" e "15 a 19 anos" apresentam porcentagem de conclusão inferior a 1%. Porém, esse resultado pode ser ocasionado pela pequena quantidade de registros para o grupo.

Tabela 8 ▶
Quantitativo de registros na dimensão faixa etária nos cursos superiores no IFPB.
Fonte: dados da pesquisa

Dimensão faixa etária	Quantidade	Concluintes	Evadidos
Menor de 15 anos	2 (0,04%)	0 (0,00%)	2 (100,00%)
15 a 19 anos	239 (5,29%)	1 (0,42%)	238 (99,58%)
20 a 24 anos	1.494 (33,10%)	458 (30,66%)	1.036 (69,34%)
25 a 29 anos	1.169 (25,90%)	445 (38,07%)	724 (61,93%)
30 a 34 anos	735 (16,28%)	213 (28,98%)	522 (71,02%)
35 a 39 anos	425 (9,42%)	108 (25,41%)	317 (74,59%)
40 a 44 anos	225 (4,98%)	69 (30,67%)	156 (69,33%)
45 a 49 anos	108 (2,39%)	35 (32,41%)	73 (67,59%)
50 a 54 anos	69 (1,53%)	23 (33,33%)	46 (66,67%)
55 a 59 anos	34 (0,75%)	17 (50,00%)	17 (50,00%)
Maior de 60 anos	14 (0,31%)	7 (50,00%)	7 (50,00%)
Total	**4.514**	**1.376 (30,48%)**	**3.138 (69,52%)**

Os gráficos da Figura 5 foram construídos a partir da análise das quantidades de registros e porcentagens de conclusão dos alunos dos cursos superiores no Brasil e no IFPB. Os gráficos apresentam uma estrutura similar quando comparados os dados nacionais e os do IFPB, com um destaque para os grupos do IFPB da faixa etária de "55 a 59 anos" e "Maior de 60 anos", que apresentam uma distância acima de 10 pontos percentuais entre os demais conjuntos.

Figura 5 ►
Análise dos cursos superiores em relação à faixa etária. (a) Quantidade por categoria no Brasil. (b) Porcentagem de concluintes no Brasil. (c) Quantidade por categoria no IFPB. (d) Porcentagem de concluintes no IFPB.
Fonte: dados da pesquisa



Os resultados decorrentes da análise quantitativa dos dados sobre faixa etária dos cursos técnicos no Brasil são mostrados na Tabela 9 (próxima página). Com exceção do grupo "Menor de 15 anos", que apresenta percentual de conclusão próximo a 30%, os demais grupos possuem um percentual superior a 40%, tendo o grupo "15 a 19 anos" um percentual de conclusão superior a 60%.

Tabela 9 ▶
Quantitativo de registros na dimensão faixa etária nos cursos técnicos no Brasil.
Fonte: dados da pesquisa

Dimensão faixa etária	Quantidade	Concluintes	Evadidos
Menor de 15 anos	133 (0,04%)	40 (30,08%)	93 (69,92%)
15 a 19 anos	121.855 (39,64%)	73.681 (60,47%)	48.174 (39,53%)
20 a 24 anos	85.603 (27,85%)	43.082 (50,33%)	42.521 (49,67%)
25 a 29 anos	36.149 (11,76%)	14.813 (40,98%)	21.336 (59,02%)
30 a 34 anos	23.661 (7,70%)	9.630 (40,70%)	14.031 (59,30%)
35 a 39 anos	16.526 (5,38%)	6.889 (41,69%)	9.637 (58,31%)
40 a 44 anos	9.982 (3,25%)	4.495 (45,03%)	5.487 (54,97%)
45 a 49 anos	6.348 (2,07%)	2.921 (46,01%)	3.427 (53,99%)
50 a 54 anos	4.054 (1,32%)	1.983 (48,91%)	2.071 (51,09%)
55 a 59 anos	1.989 (0,65%)	952 (47,86%)	1.037 (52,14%)
Maior de 60 anos	1.078 (0,35%)	490 (45,45%)	588 (54,55%)
Total	**307.378**	**158.976 (51,72%)**	**148.402 (48,28%)**

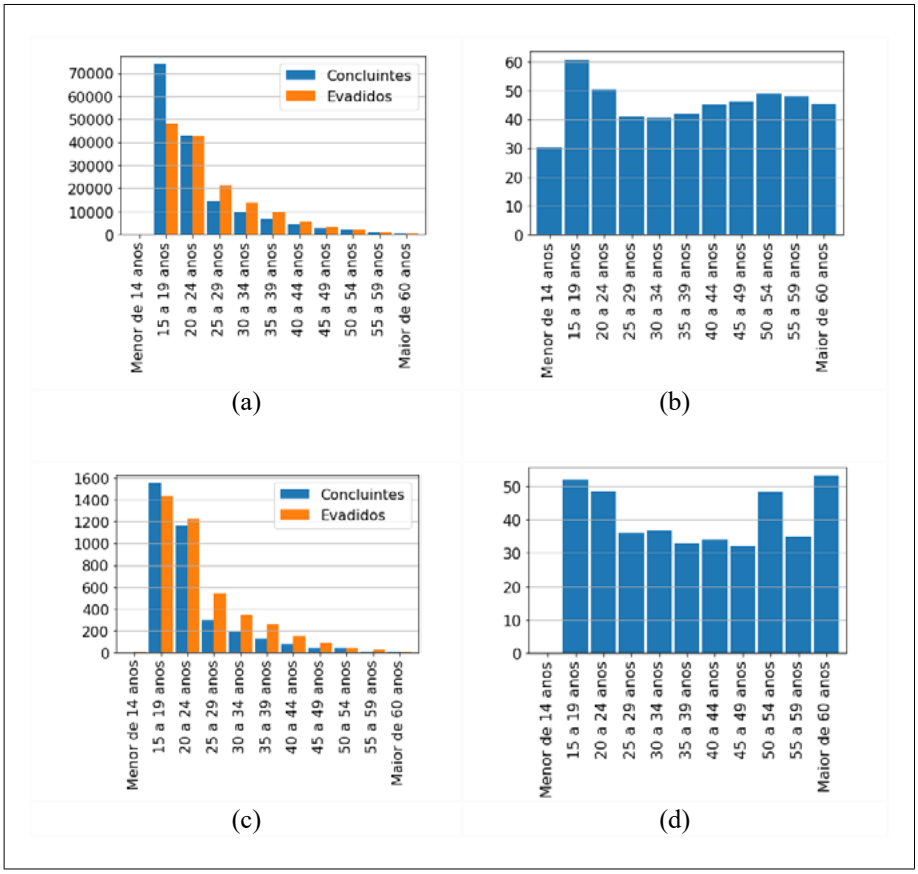
A análise quantitativa dos dados sobre faixa etária no DW para o conjunto de dados dos cursos técnicos no IFPB é mostrada na Tabela 10. Com os resultados, é possível observar que apenas os grupos “15 a 19 anos” e “Maior de 60 anos” apresentam um percentual de conclusão superior a 50%.

Tabela 10 ▶
Quantitativo de registros na dimensão faixa etária nos cursos técnicos no IFPB.
Fonte: dados da pesquisa

Dimensão faixa etária	Quantidade	Concluintes	Evadidos
Menor de 15 anos	7 (0,09%)	0 (0,00%)	7 (100,00%)
15 a 19 anos	2.980 (39,02%)	1.548 (51,95%)	1.432 (48,05%)
20 a 24 anos	2.393 (31,33%)	1.164 (48,64%)	1.229 (51,36%)
25 a 29 anos	842 (11,02%)	304 (36,10%)	538 (63,90%)
30 a 34 anos	536 (7,02%)	196 (36,57%)	340 (63,43%)
35 a 39 anos	385 (5,04%)	127 (32,99%)	258 (67,01%)
40 a 44 anos	226 (2,96%)	77 (34,07%)	149 (65,93%)
45 a 49 anos	125 (1,64%)	40 (32,00%)	85 (68,00%)
50 a 54 anos	87 (1,14%)	42 (48,28%)	45 (51,72%)
55 a 59 anos	40 (0,52%)	14 (35,00%)	26 (65,00%)
Maior de 60 anos	17 (0,22%)	9 (52,94%)	8 (47,06%)
Total	**7.638**	**3.521 (46,10%)**	**4.117 (53,90%)**

A Figura 6 apresenta os gráficos construídos a partir da análise das quantidades de registros e porcentagens de conclusão dos alunos dos cursos técnicos no Brasil e no IFPB. O destaque, quando comparados os resultados, é observado no crescimento no grupo “Maior de 60 anos” para os dados do IFPB em relação aos dados nacionais.

Figura 6 ▶
Análise dos cursos técnicos em relação à faixa etária. (a) Quantidade por categoria no Brasil. (b) Porcentagem de concluintes no Brasil. (c) Quantidade por categoria no IFPB. (d) Porcentagem de concluintes no IFPB.
Fonte: dados da pesquisa



4.4 Análise por renda familiar

Os resultados a seguir são decorrentes da análise dos dados sobre a dimensão que registra o valor da Renda Familiar por Pessoa (RFP) em relação ao valor do salário mínimo declarado pelo aluno. A Tabela 11 mostra os resultados decorrentes da análise quantitativa sobre renda dos alunos de cursos superiores no Brasil. Nela, percebe-se que, entre os alunos nas faixas até 1,5 salário mínimo, o percentual de conclusão foi próximo de 30%. Já entre os alunos cuja RFP é maior do que 1,5 salário mínimo, o percentual de conclusão foi próximo de 35%.

Tabela 11 ▶
Quantitativo de registros na dimensão renda familiar nos cursos superiores no Brasil.
Fonte: dados da pesquisa

Dimensão renda familiar	Quantidade	Concluintes	Evadidos
0 < RFP <= 0,5	14.821 (13,39%)	4.542 (30,65%)	10.279 (69,35%)
0,5 < RFP <= 1,0	15.676 (14,16%)	4.956 (31,62%)	10.720 (68,38%)
1,0 < RFP <= 1,5	10.584 (9,56%)	3.233 (30,55%)	7.351 (69,45%)
1,5 < RFP <= 2,5	8.465 (7,65%)	2.998 (35,42%)	5.467 (64,58%)
2,5 < RFP <= 3,5	3.523 (3,18%)	1.279 (36,30%)	2.244 (63,70%)
RFP > 3,5	4.433 (4,00%)	1.591 (35,89%)	2.842 (64,11%)
Não declarada	53.215 (48,06%)	15.591 (29,30%)	37.624 (70,70%)
Total	**110.717**	**34.190 (30,88%)**	**76.527 (69,12%)**

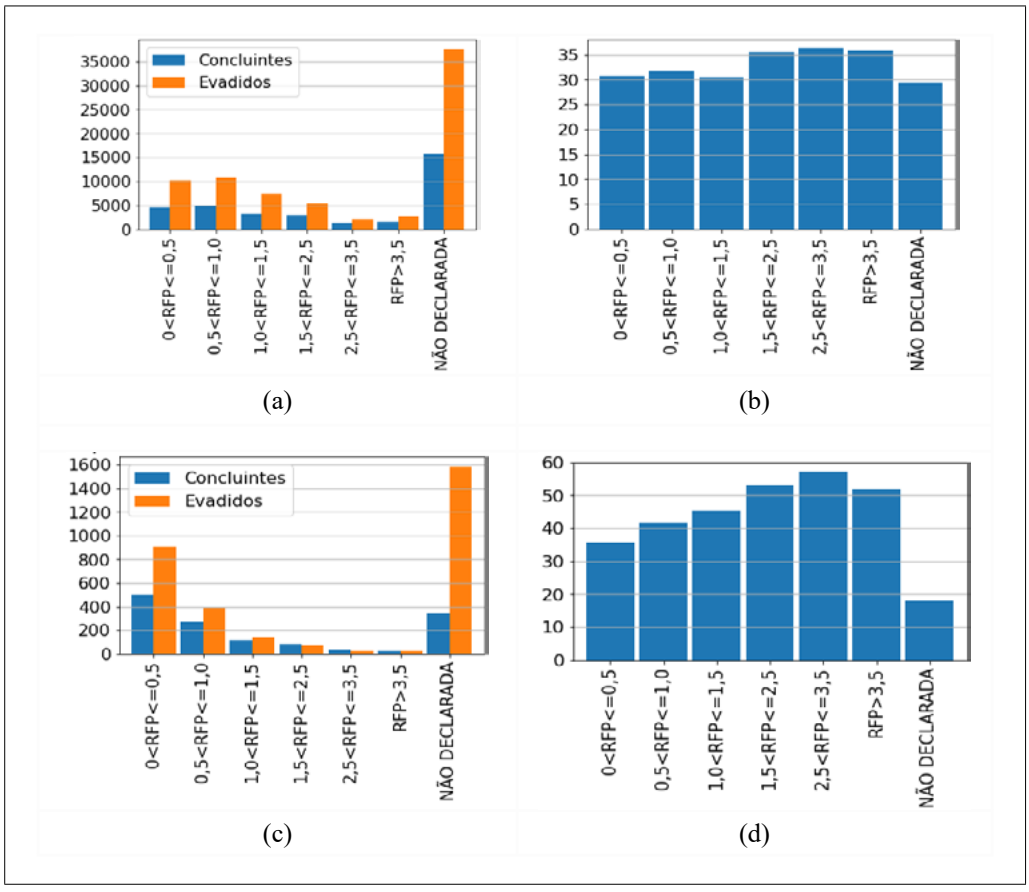
A Tabela 12 mostra os resultados decorrentes da análise quantitativa dos dados sobre renda familiar no DW para o conjunto de dados dos cursos superiores no IFPB. Nela, é possível observar o aumento para o grupo “2,5 < RFP <= 3,5”, que contabiliza um percentual de +20,84 pontos percentuais de concluintes para o IFPB quando comparado aos dados nacionais.

Tabela 12 ▶
Quantitativo de registros na dimensão renda familiar nos cursos superiores no IFPB.
Fonte: dados da pesquisa

Dimensão renda familiar	Quantidade	Concluintes	Evadidos
0 < RFP <= 0,5	1.402 (31,06%)	500 (35,66%)	902 (64,34%)
0,5 < RFP <= 1,0	658 (14,58%)	273 (41,49%)	385 (58,51%)
1,0 < RFP <= 1,5	266 (5,89%)	120 (45,11%)	146 (54,89%)
1,5 < RFP <= 2,5	149 (3,30%)	79 (53,02%)	70 (46,98%)
2,5 < RFP <= 3,5	56 (1,24%)	32 (57,14%)	24 (42,86%)
RFP > 3,5	50 (1,11%)	26 (52,00%)	24 (48,00%)
Não declarada	1.933 (42,82%)	346 (17,90%)	1.587 (82,10%)
Total	**4.514**	**1.376 (30,48%)**	**3.138 (69,52%)**

A Figura 7 apresenta os gráficos construídos a partir da análise das quantidades de registros e percentuais de conclusão dos alunos dos cursos superiores no Brasil e no IFPB. Quando comparados os dois gráficos, observa-se um crescimento nas porcentagens de conclusão conforme o crescimento da renda familiar, de forma que, para o conjunto do IFPB, esse crescimento aparenta ser mais acentuado.

Figura 7 ▶
Análise dos cursos superiores em relação à renda familiar.
(a) Quantidade por categoria no Brasil.
(b) Porcentagem de concluintes no Brasil.
(c) Quantidade por categoria no IFPB.
(d) Porcentagem de concluintes no IFPB.
Fonte: dados da pesquisa



A Tabela 13 mostra os resultados decorrentes da análise quantitativa dos dados sobre renda familiar no DW para o conjunto de dados dos cursos técnicos no Brasil. É possível observar um crescimento percentual de conclusão dos alunos conforme há o crescimento da renda familiar, de modo que a diferença entre os grupos "0 < RFP <= 0,5" e "RFP > 3,5" é de 19,66%.

Tabela 13 ▶
Quantitativo de registros na dimensão renda familiar nos cursos técnicos no Brasil.
Fonte: dados da pesquisa

Dimensão renda familiar	Quantidade	Concluintes	Evadidos
0 < RFP <= 0,5	58.980 (19,19%)	29.912 (50,72%)	29.068 (49,28%)
0,5 < RFP <= 1,0	48.953 (15,93%)	26.204 (53,53%)	22.749 (46,47%)
1,0 < RFP <= 1,5	28.010 (9,11%)	15.152 (54,10%)	12.858 (45,91%)
1,5 < RFP <= 2,5	18.064 (5,88%)	10.708 (59,28%)	7.356 (40,72%)
2,5 < RFP <= 3,5	6.977 (2,27%)	4.172 (59,80%)	2.805 (40,20%)
RFP > 3,5	8.123 (2,64%)	5.435 (66,91%)	2.688 (33,09%)
Não declarada	138.271 (44,98%)	67.393 (48,74%)	70.878 (51,26%)
Total	**307.378**	**158.976 (51,72%)**	**148.402 (48,28%)**

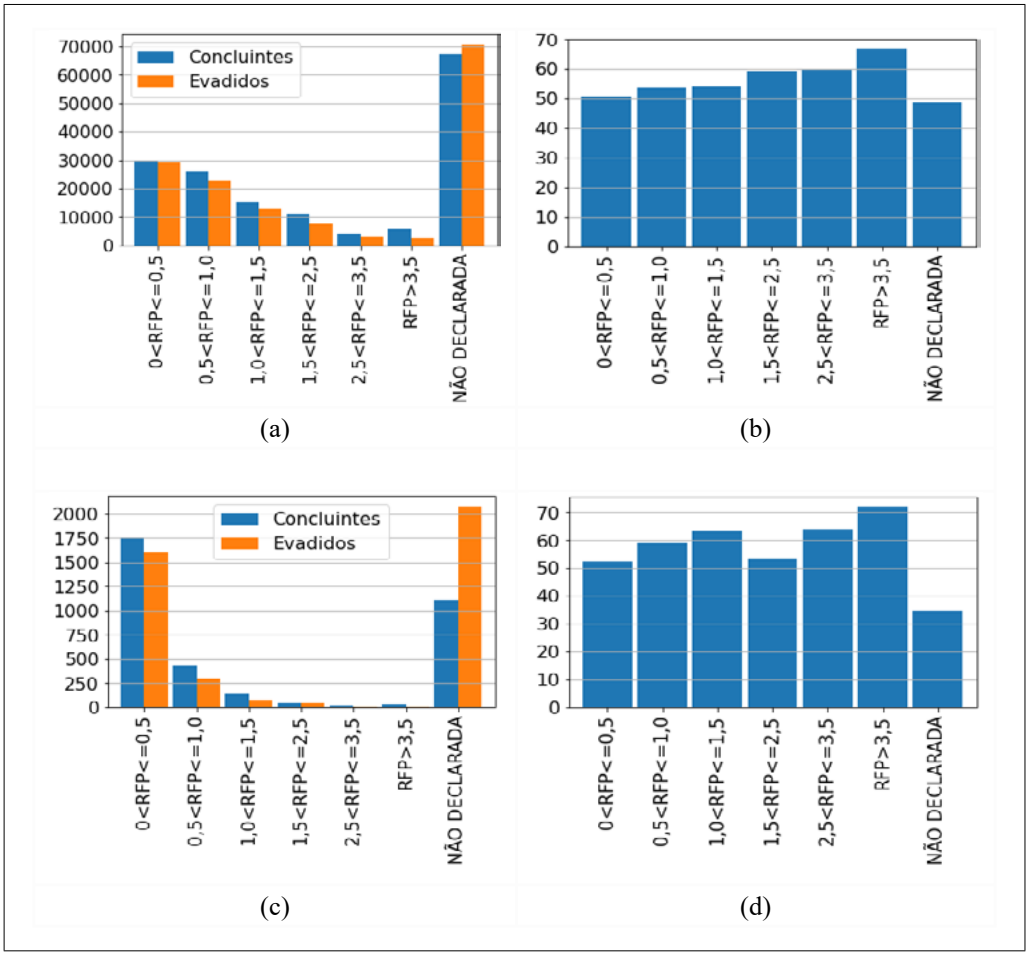
A Tabela 14 mostra os resultados decorrentes da análise quantitativa dos dados sobre renda familiar no DW para o conjunto de dados dos cursos técnicos no IFPB. Nos dados do IFPB há também a ocorrência do crescimento da porcentagem de conclusão entre os grupos "0 < RFP <= 0,5" e "RFP > 3,5", com uma diferença de +19,6 pontos percentuais. Porém, existe uma redução no grupo "1,5 < RFP <= 2,5", que quase se equipara ao primeiro grupo, com apenas 0,58% a mais quando comparado à faixa "0 < RFP <= 0,5".

Tabela 14 ▶
Quantitativo de registros na dimensão renda familiar nos cursos técnicos no IFPB.
Fonte: dados da pesquisa

Dimensão renda familiar	Quantidade	Concluintes	Evadidos
0 < RFP <= 0,5	3.357 (43,95%)	1.760 (52,43%)	1.597 (47,57%)
0,5 < RFP <= 1,0	726 (9,51%)	427 (58,82%)	299 (41,18%)
1,0 < RFP <= 1,5	218 (2,85%)	138 (63,30%)	80 (36,70%)
1,5 < RFP <= 2,5	83 (1,09%)	44 (53,01%)	39 (46,99%)
2,5 < RFP <= 3,5	25 (0,33%)	16 (64,00%)	9 (36,00%)
RFP > 3,5	43 (0,56%)	31 (72,09%)	12 (27,91%)
Não declarada	3.186 (41,71%)	1.105 (34,68%)	2.081 (65,32%)
Total	**7.638**	**3.521 (46,10%)**	**4.117 (53,90%)**

A Figura 8 apresenta os gráficos construídos a partir da análise da quantidade de registros e percentuais de conclusão dos alunos dos cursos técnicos no Brasil e no IFPB. Comparados os dois gráficos, observa-se um crescimento nas porcentagens de conclusão conforme o crescimento da renda familiar, semelhante ao observado para os cursos superiores. Porém, é observada uma irregularidade no crescimento para o conjunto de dados do IFPB.

Figura 8 ▶
Análise dos cursos técnicos em relação à renda familiar. (a) Quantidade por categoria no Brasil. (b) Porcentagem de concluintes no Brasil. (c) Quantidade por categoria no IFPB. (d) Porcentagem de concluintes no IFPB.
Fonte: dados da pesquisa



4.5 Dados de gênero/sexo

Os resultados desta subseção são decorrentes da análise sobre o gênero/sexo nos dados dos alunos no DW. A Tabela 15 mostra os resultados decorrentes da análise quantitativa para o conjunto de dados dos cursos superiores no Brasil. Nela, é possível observar que, apesar da maior quantidade de alunos do sexo masculino em relação ao sexo feminino, com +19,26 pontos percentuais, as alunas têm uma porcentagem de conclusão superior aos alunos em 7,8 pontos percentuais.

Tabela 15 ▶
Quantitativo de registros na dimensão gênero/sexo nos cursos superiores no Brasil.
Fonte: dados da pesquisa

Dimensão gênero/sexo	Quantidade	Concluintes	Evadidos
Masculino	66.026 (59,63%)	18.312 (27,73%)	47.714 (72,27%)
Feminino	44.691 (40,37%)	15.878 (35,53%)	28.813 (64,47%)
Total	**110.717**	**34.190 (30,88%)**	**76.527 (69,12%)**

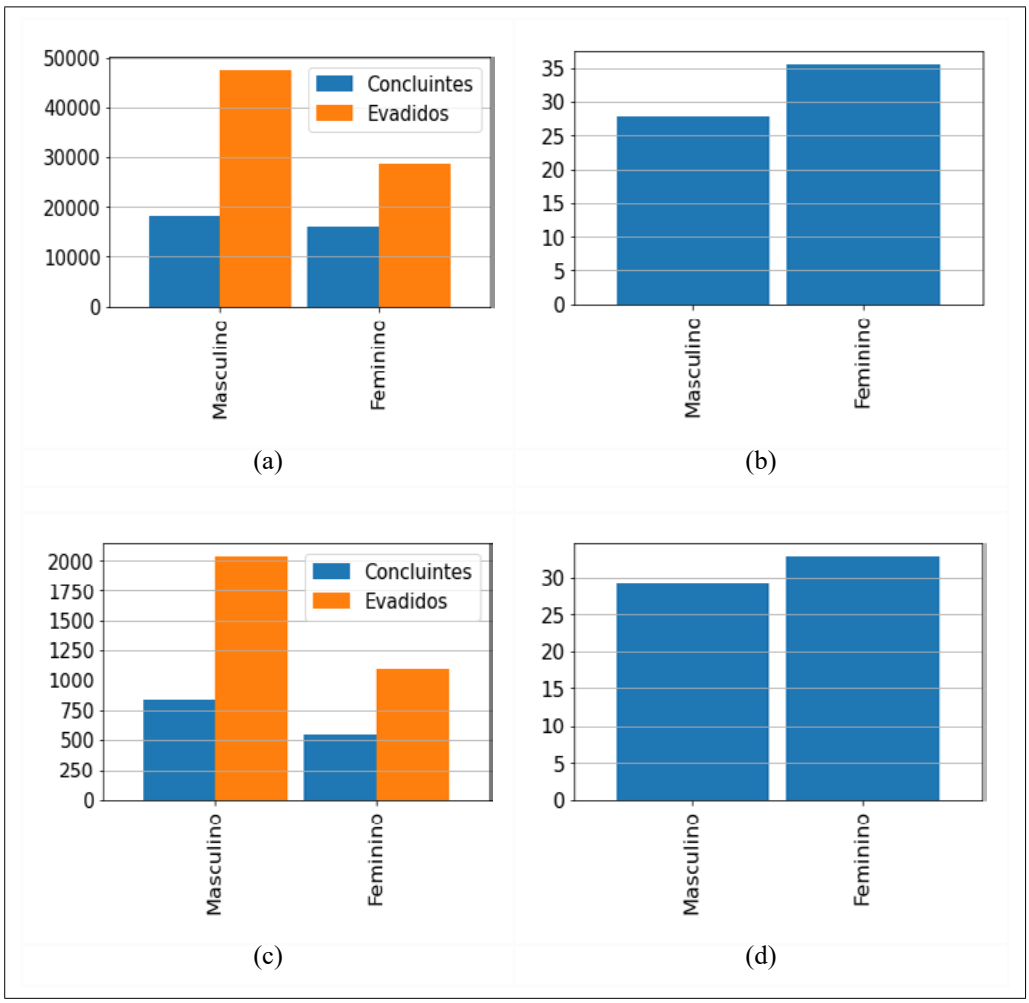
A Tabela 16 mostra os resultados decorrentes da análise quantitativa dos dados sobre gênero/sexo no DW para o conjunto de dados dos cursos superiores no IFPB. Semelhante ao que ocorre no cenário nacional, existe uma maior quantidade de alunos do sexo masculino, com uma diferença de 27,56 pontos percentuais, entretanto os alunos do sexo feminino possuem um percentual de conclusão 3,6 pontos percentuais superior quando comparados com os alunos do sexo masculino. Isso representa uma redução quando comparado ao cenário nacional.

Tabela 16 ▶
Quantitativo de registros na dimensão gênero/sexo nos cursos superiores no IFPB.
Fonte: dados da pesquisa

Dimensão gênero/sexo	Quantidade	Concluintes	Evadidos
Masculino	2.879 (63,78%)	840 (29,18%)	2.039 (70,82%)
Feminino	1.635 (36,22%)	536 (32,78%)	1.099 (67,22%)
Total	**4.514**	**1.376 (30,48%)**	**3.138 (69,52%)**

A Figura 9 apresenta os gráficos construídos a partir da análise das quantidades de registros e porcentagens de conclusão dos alunos dos cursos superiores no Brasil e no IFPB. Nos dois gráficos é possível observar que, apesar da diferença da quantidade de alunos entre os dois sexos, a evasão do sexo masculino é maior proporcionalmente que a do sexo feminino, de maneira que quase se equiparam as quantidades numéricas de conclusão entre os sexos masculino e feminino.

Figura 9 ▶
Análise dos cursos superiores em relação ao gênero/sexo. (a) Quantidade por categoria no Brasil. (b) Porcentagem de concluintes no Brasil. (c) Quantidade por categoria no IFPB. (d) Porcentagem de concluintes no IFPB.
Fonte: dados da pesquisa



A Tabela 17 (próxima página) mostra os resultados decorrentes da análise quantitativa dos dados sobre gênero/sexo no DW para o conjunto de dados dos cursos técnicos no Brasil. Nela, é possível observar que o cenário é semelhante ao que ocorre com os cursos superiores: uma maior quantidade de alunos do sexo masculino, com uma diferença de 7,54 pontos percentuais, porém há um maior percentual de conclusão dos alunos do sexo feminino, com uma diferença de 3,08 pontos percentuais.

Tabela 17 ▶
Quantitativo de registros na dimensão gênero/sexo nos cursos técnicos no Brasil.
Fonte: dados da pesquisa

Dimensão gênero/sexo	Quantidade	Concluintes	Evadidos
Masculino	165.286 (53,77%)	83.131 (50,30%)	82.155 (49,70%)
Feminino	142.092 (46,23%)	75.845 (53,38%)	66.247 (46,62%)
Total	**307.378**	**158.976 (51,72%)**	**148.402 (48,28%)**

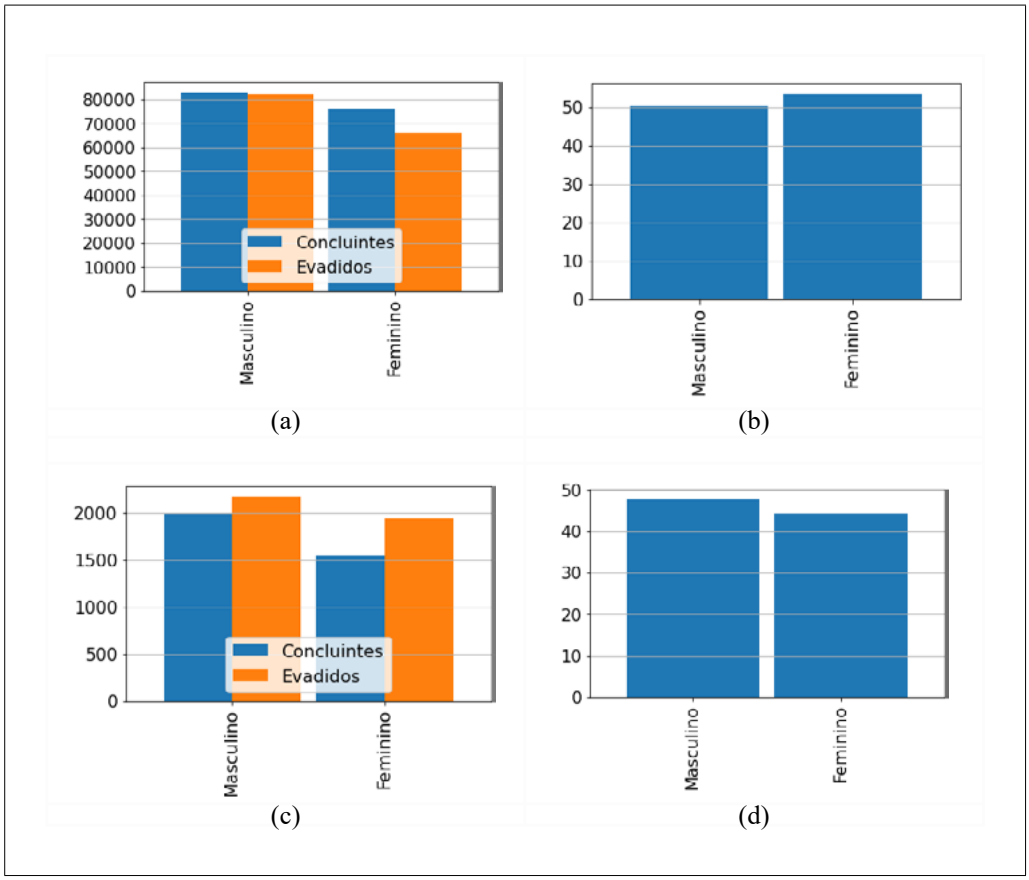
A Tabela 18 mostra os resultados decorrentes da análise quantitativa dos dados sobre gênero/sexo no DW para o conjunto de dados dos cursos técnicos no IFPB. Nesses dados, há uma quantidade maior de alunos do sexo masculino, +8,88 pontos percentuais, quando comparado ao sexo feminino. Mesmo existindo semelhança na diferença de proporção entre os sexos masculino e feminino, existe uma inversão nos valores da porcentagem de conclusão quando comparado aos demais cenários. O sexo masculino possui uma diferença superior de 3,39 pontos percentuais em comparação ao sexo feminino para a porcentagem de conclusão.

Tabela 18 ▶
Quantitativo de registros na dimensão gênero/sexo nos cursos técnicos no IFPB.
Fonte: dados da pesquisa

Dimensão gênero/sexo	Quantidade	Concluintes	Evadidos
Masculino	4.158 (54,44%)	1.981 (47,64%)	2.177 (52,36%)
Feminino	3.480 (45,56%)	1.540 (44,25%)	1.940 (55,75%)
Total	**7.638**	**3.521 (46,10%)**	**4.117 (53,90%)**

A Figura 10 apresenta os gráficos construídos a partir da análise das quantidades de registros e porcentagens de conclusão dos alunos dos cursos técnicos no Brasil e no IFPB. Nesses gráficos é possível observar a inversão dos percentuais de conclusão entre os dados nacionais e os do IFPB.

Figura 10 ▶
Análise dos cursos técnicos em relação ao gênero/sexo. (a) Quantidade por categoria no Brasil. (b) Porcentagem de concluintes no Brasil. (c) Quantidade por categoria no IFPB. (d) Porcentagem de concluintes no IFPB.
Fonte: dados da pesquisa



5 Conclusão

Os dados da Plataforma Nilo Peçanha permitem uma análise, em diversas perspectivas, dos dados oriundos dos cursos de instituições de ensino federais no Brasil. A utilização desses dados possibilita a construção de um DW, com o intuito de proporcionar a recuperação dos dados e uma análise detalhada sobre a evasão escolar em diversos cenários.

A análise dos dados demonstra que alguns atributos se destacam quando analisada a porcentagem de conclusão dos alunos, sendo possível constatar que faixa etária, renda per capita familiar e etnia apresentam relevância ao se observar os agrupamentos dos dados quanto à evasão escolar. A utilização do DW para gerar consultas direcionadas a características específicas facilita a análise dos dados focando em espectros específicos como o da evasão escolar, além de possibilitar comparações entre instituições de ensino e o cenário nacional. A aplicação de consultas em um DW viabiliza a adaptação das consultas, de forma a incluir ou remover atributos conforme a necessidade e o foco da pesquisa.

Como próximas etapas, a partir dos resultados obtidos neste trabalho, é possível expandir a análise para incluir mais dados provenientes de sistemas de gerenciamento acadêmico, como forma de identificar um conjunto maior de atributos mais relevantes relacionados à evasão escolar.

Financiamento

Esta pesquisa foi financiada pelo IFPB por meio da Chamada Interconecta IFPB nº 2/2021 – Apoio a projetos de Pesquisa, Inovação, Desenvolvimento Tecnológico e Social.

Conflito de interesses

Os autores declaram não haver conflito de interesses.

Referências

BAKER, R.; ISOTANI, S.; CARVALHO, A. Mineração de dados educacionais: oportunidades para o Brasil. **Revista Brasileira de Informática na Educação**, v. 19, n. 2, p. 3-13, 2011. DOI: http://dx.doi.org/10.5753/rbie.2011.19.02.03.

BRITO, D. M.; PASCOAL, T. A.; ARAÚJO, J. G. G. O.; LEMOS, M. O.; RÊGO, T. G. Identificação de estudantes do primeiro semestre com risco de evasão através de técnicas de Data Mining. *In*: XX CONGRESSO INTERNACIONAL DE INFORMÁTICA EDUCATIVA (TISE 2015), 20., 2015, Santiago. **Nuevas Ideas en Informática Educativa**: v. 11. Santiago: Universidad de Chile, 2015. Disponível em: http://www.tise.cl/volumen11/TISE2015/459-463.pdf. Acesso em: 12 nov. 2021.

CARPENTER, J.; HEWITT, E. **Cassandra**: The definitive guide. 3. ed. Sebastopol: O'Reilly Media, 2020.

CASTRO, L. N.; FERRARI, D. G. **Introdução à mineração de dados**: conceitos básicos, algoritmos e aplicações. São Paulo: Saraiva, 2016.

DORE, R.; LÜSCHER, A. Z. Permanência e evasão na educação técnica de nível médio em Minas Gerais. **Cadernos de pesquisa**, v. 41, n. 144, p. 770-789, 2011. DOI: https://doi.org/10.1590/S0100-15742011000300007.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI Magazine**, v. 17, n. 3, p. 37-54, 1996. DOI: https://doi.org/10.1609/aimag.v17i3.1230.

FERREIRA, J. T. A.; MIRANDA, M.; ABELHA, A.; MACHADO, J. M. O processo ETL em sistemas *Data Warehouse*. *In*: SIMPÓSIO DE INFORMÁTICA, 2., 2010, Braga. **Actas** [...]. Braga: Universidade do Minho, 2010. p. 757-765. Disponível em: http://repositorium.sdum.uminho.pt/handle/1822/11435. Acesso em: 3 nov. 2022.

INEP – INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA. **Sinopse Estatística da Educação Superior**. 2020. Disponível em https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados. Acesso em: 22 nov. 2020.

INMON, W. H. **Building the data warehouse**. 4th ed. New York: Wiley, 2002.

KABASHI, Q.; SHABANI, I.; CAKA, N. Analysis of the student dropout rate at the Faculty of Electrical and Computer Engineering of the University of Prishtina, Kosovo, From 2001 to 2015. **IEEE Access**, v. 10, p. 68126-68137, 2022. DOI: https://doi.org/10.1109/ACCESS.2022.3185620.

KAUR, P.; SINGH, M.; JOSAN, G. S. Classification and prediction based data mining algorithms to predict slow learners in education sector. **Procedia Computer Science**, v. 57, p. 500-508, 2015. DOI: https://doi.org/10.1016/j.procs.2015.07.372.

KEMCZINSKI, A.; CIDRAL, A.; CASTRO, J. E. E.; FIOD NETO, M. Como obter vantagem competitiva utilizando business intelligence? **Produção On Line**, v. 3, n. 2, 2003. DOI: https://doi.org/10.14488/1676-1901.v3i2.626.

KIMBALL, R.; ROSS, M. **The data warehouse toolkit**: the definitive guide to dimensional modeling. 3rd ed. Indianapolis: Wiley, 2013.

MNYAWAMI, Y. N.; MAZIKU, H. H.; MUSHI, J. C. Enhanced model for predicting student dropouts in developing countries using automated machine learning approach: a case of tanzanian's secondary schools. **Applied Artificial Intelligence**, v. 36, n. 1, p. 433-451, 2022. DOI: https://doi.org/10.1080/08839514.2022.2071406.

MOLINER, L.; ALEGRE, F.; LORENZO-VALENTIN, G. The COVID-19 Pandemic's Impact on 9th Grade Students' Mathematics Achievement. **European Journal of Educational Research**, v. 11, n. 2, p. 835-845, 2022. DOI: https://doi.org/10.12973/eu-jer.11.2.835.

PRIM, A. L.; FÁVERO, J. D. Motivos da evasão escolar nos cursos de ensino superior de uma faculdade na cidade de Blumenau. **E-Tech: Tecnologias para**

Competitividade Industrial, Florianópolis, 3. ed. especial, p. 53-72, 2013. DOI: https://doi.org/10.18624/e-tech.v0i0.382.

ROMERO, C.; VENTURA, S.; PECHENIZKY, M.; BAKER, R. S. F. **Handbook of educational data mining**. New York: Taylor & Francis, 2010.

SAKKA, A.; BIMONTE, S.; PINET, F.; SAUTOT, L. Volunteer data warehouse: state of the art. **International Journal of Data Warehousing and Mining (IJDWM)**, v. 17, n. 3, p. 1-21, 2021. DOI: https://dx.doi.org/10.4018/IJDWM.2021070101.

SANTOS, F. F. P.; SIMON, L. M.; PINTO, N. G. M. Retenção e evasão escolar em um Instituto Federal de Educação, Ciência e Tecnologia. **Revista Científica da AJES**, v. 9, n. 18, p. 186-196, 2020. Disponível em: https://revista.ajes.edu.br/index.php/rca/article/view/334. Acesso em: 3 nov. 2022.

SILVA, T. H. O.; MENDONÇA, F. M. Análise de evasão dos cursos de licenciatura de uma instituição de ensino pertencente a rede federal de educação, profissional, científica e tecnológica. Curitiba. **Brazilian Journal of Development**, v. 7, n. 8, p. 80739-80751, 2021. DOI: https://doi.org/10.34117/bjdv7n8-338.

SINGH, H. P.; ALHULAIL, H. N. Predicting student-teachers dropout risk and early identification: a four-step logistic regression approach. **IEEE Access**, v. 10, p. 6470-6482, 2022. DOI: http://doi.org/10.1109/ACCESS.2022.3141992.

VEIGA, C. R.; BERGIANTE, N. Fatores predominantes da evasão escolar no ensino médio profissional: uma revisão de literatura. *In*: CONGRESSO NACIONAL DE EXCELÊNCIA EM GESTÃO, 12.; INOVARSE, 3., 2016, Rio de Janeiro. **Anais** [...]. Rio de Janeiro: CNEG, 2016.

XU, C.; ZHU, G.; YE, J.; SHU, J. Educational data mining: dropout prediction in XuetangX MOOCs. **Neural Processing Letters**, v. 54, p. 2885-2900, 2022. DOI: https://doi.org/10.1007/s11063-022-10745-5.