

**INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DA PARAÍBA
CAMPUS CAJAZEIRAS
CURSO SUPERIOR DE TECNOLOGIA EM ANÁLISE E DESENVOLVIMENTO DE
SISTEMAS**

**ANALISANDO A PNP: O DESENVOLVIMENTO DE UM DATA
WAREHOUSE PARA MINERAÇÃO DE DADOS SOBRE EVASÃO NA
REDE FEDERAL**

MICHEL GABRIEL MOREIRA

**Cajazeiras
2022**

MICHEL GABRIEL MOREIRA

**ANALISANDO A PNP: O DESENVOLVIMENTO DE UM DATA WAREHOUSE PARA
MINERAÇÃO DE DADOS SOBRE EVASÃO NA REDE FEDERAL**

Trabalho de Conclusão de Curso apresentado junto ao Curso Superior de Tecnologia em Análise e Desenvolvimento de Sistemas do Instituto Federal de Educação, Ciência e Tecnologia da Paraíba - Campus Cajazeiras, como requisito à obtenção do título de Tecnólogo em Análise e Desenvolvimento de Sistemas.

Orientador

Prof. Me. Francisco Paulo de Freitas Neto.

**Cajazeiras
2022**

IFPB / Campus Cajazeiras
Coordenação de Biblioteca
Biblioteca Prof. Ribamar da Silva
Catalogação na fonte: Suellen Conceição Ribeiro CRB-2218

M838a Moreira, Michel Gabriel

Analisando a PNP: o desenvolvimento de um data warehouse para mineração de dados sobre evasão na rede federal / Michel Gabriel Moreira. – Cajazeiras/PB: IFPB, 2022.

61f.:il.

Trabalho de Conclusão de Curso (Tecnólogo em Análise e Desenvolvimento de Sistemas) - Instituto Federal de Educação, Ciência e Tecnologia da Paraíba-IFPB, Campus Cajazeiras. Cajazeiras, 2022.

Orientador(a): Prof. Me. Francisco Paulo de Freitas Neto.

1. Informática. 2. Evasão. 3. Plataforma Nilo Peçanha. 4. Rede Federal. 5 *Data Warehouse*.

I. Moreira, Michel Gabriel. II. Título

CDU: 004 M838a

ATA 32/2022 - CADS/UNINFO/DDE/DG/CZ/REITORIA/IFPB

ATA DE DEFESA DO TRABALHO DE CONCLUSÃO DE CURSO (TCC)
CURSO: ANÁLISE E DESENVOLVIMENTO DE SISTEMAS (ADS)

Às 14h00 do dia 05 do mês de MAIO do ano de 2022, o(a) aluno(a) **MICHEL GABRIEL MOREIRA**, matrícula **201912010005**, apresentou, como parte dos requisitos para obtenção do título de Tecnólogo em Análise e Desenvolvimento de Sistemas, seu trabalho de conclusão de curso, tendo como título "**ANALISANDO A PNP: O DESENVOLVIMENTO DE UM DATA WAREHOUSE PARA MINERAÇÃO DE DADOS SOBRE EVASÃO NA REDE FEDERAL**". Constituíram a banca examinadora os professores **Francisco Paulo de Freitas Neto** (orientador), **Fábio Gomes de Andrade** (examinador) e **Ricardo de Sousa Job** (examinador).

Após a apresentação e as observações dos membros da Banca Examinadora, ficou definido que o trabalho foi considerado **APROVADO** com nota **90**, com a condição de que o (a) aluno (a) entregue, no prazo máximo de 30 dias, a versão final do trabalho com as correções sugeridas pelos membros da banca examinadora. Eu, **FÁBIO ABRANTES DINIZ**, Coordenador do Curso Superior de Tecnologia em Análise e Desenvolvimento de Sistemas, lavrei a presente ata, que segue assinada digitalmente por mim e pelos membros da banca examinadora.

Cajazeiras, 11 de maio de 2022.

Documento assinado eletronicamente por:

- Francisco Paulo de Freitas Neto, PROFESSOR ENS BASICO TECN TECNOLOGICO, em 11/05/2022 17:47:44.
- Michel Gabriel Moreira, ALUNO (201912010005) DE TECNOLOGIA EM ANÁLISE E DESENVOLVIMENTO DE SISTEMAS - CAJAZEIRAS, em 12/05/2022 13:55:47.
- Fabio Gomes de Andrade, PROFESSOR ENS BASICO TECN TECNOLOGICO, em 12/05/2022 16:10:42.
- Ricardo de Sousa Job, PROFESSOR ENS BASICO TECN TECNOLOGICO, em 12/05/2022 18:38:33.

Este documento foi emitido pelo SUAP em 11/05/2022. Para comprovar sua autenticidade, faça a leitura do QRCode ao lado ou acesse <https://suap.ifpb.edu.br/autenticar-documento/> e forneça os dados abaixo:

Código Verificador: 293469

Código de Autenticação: f0d63cf023



Dedico este trabalho para Rosalina e Jonathas, pessoas fundamentais para que ele se tornasse realidade.

RESUMO

A evasão escolar é um problema crônico no Brasil, problema este que necessita ser estudado para que sejam encontradas soluções efetivas para tentar impedir que os alunos abandonem os estudos. A Plataforma Nilo Peçanha é responsável por agrupar dados educacionais da rede federal brasileira e divulgá-los. A análise dessas informações se torna essencial para encontrar padrões e características que possam ajudar os gestores das instituições de ensino a agirem com mais exatidão e sucesso no combate da evasão. Esse trabalho realiza a organização dos dados da Plataforma Nilo Peçanha em um *Data Warehouse*. Esses dados foram utilizados em algoritmos de classificação como árvores de decisão e redes neurais artificiais. Foram identificadas características relevantes na classificação de perfis evadidos com uma acurácia de até 80%.

Palavras-chave: Evasão. Plataforma Nilo Peçanha. *Data Warehouse*. Modelagem Dimensional. Mineração de dados. Árvores de Decisão. Redes Neurais.

ABSTRACT

School dropout is a chronic problem in Brazil, a problem that needs to be studied so that effective solutions are found to try to prevent students from dropping out of their studies. The Nilo Peçanha Platform is responsible for clustering this data from the Brazilian federal network and disseminating it. The analysis of said information becomes essential to find patterns and characteristics that can help the managers of educational institutions to act more accurately and successfully in combating dropout. This work carried out a reorganization of the platform's data into a Data Warehouse. Then the data was utilized with classification algorithms such as decision trees and artificial neural networks. Relevant characteristics in the identification of evaded profiles were found with an accuracy of up to 80%.

Keywords: School Dropout. Nilo Peçanha Platform. Data Warehouse. Dimensional Modeling. Data Mining. Decision Trees. Neural Networks.

LISTA DE FIGURAS

Figura 1 – Exemplo de esquema estrela	20
Figura 2 – Exemplo de esquema constelação	21
Figura 3 – Exemplo de esquema floco de neve	22
Figura 4 – Nós e Ramificações das Árvores Decisórias	28
Figura 5 – Organização em camadas	29
Figura 6 – Modelo dimensional - Esquema estrela	31
Figura 7 – Arquitetura do trabalho	33
Figura 8 – Pré-processamento do <i>dataframe</i>	36
Figura 9 – Árvore de decisão com os dados de carga horária mínima e data de fim previsto	43
Figura 10 – Árvore de decisão com dados de renda familiar	44
Figura 11 – Árvore de decisão com dados de sexo e raça	45

LISTA DE QUADROS

Quadro 1 – Parâmetros escolhidos e seus possíveis valores para Árvore de Decisão	38
Quadro 2 – Parâmetros escolhidos e seus possíveis valores para Rede Neural	40
Quadro 3 – Parâmetros com melhores performances para as árvores de decisão	42
Quadro 4 – Média dos resultados para as árvores de decisão	43
Quadro 5 – Parâmetros com melhores performances para as redes neurais artificiais	46
Quadro 6 – Média dos resultados para as redes neurais artificiais	46

LISTA DE CÓDIGOS

Algoritmo 1 – Pseudocódigo da função COPY	34
Algoritmo 2 – Pseudocódigo dos laços de repetição para execução das árvores de decisão	39
Algoritmo 3 – Pseudocódigo do laço de repetição para execução das redes neurais artificiais	41

LISTA DE ABREVIATURAS E SIGLAS

ADS	Análise e Desenvolvimento de Sistemas
API	Application Programming Interface
CSV	Comma-separated values
DW	Data Warehouse
ECCD	Extrair, Limpar, Transformar e Entregar
ETL	Extrac, Transform and Load
UNICEF	Fundo das Nações Unidas para a Infância
IES	Instituição de Ensino Superior
INESC	Instituto de Estudos Socioeconômicos
IFPB	Instituto Federal de Educação, Ciência e Tecnologia da Paraíba
JSON	JavaScript Object Notation
KDD	Knowledge Discovery in Databases
NBR	Norma Brasileira
PNP	Plataforma Nilo Peçanha
REST	Representational State Transfer
TCC	Trabalho de Conclusão do Curso

SUMÁRIO

1	INTRODUÇÃO	12
1.1	Motivação	13
1.2	Objetivo geral	14
1.3	Objetivos específicos	14
1.4	Trabalhos Relacionados	14
1.5	Metodologia	15
1.6	Organização do Documento	16
2	REFERENCIAL TEÓRICO	17
2.1	Plataforma Nilo Peçanha	17
2.2	Modelagem Dimensional	18
2.2.1	Elementos dos modelos dimensionais	19
2.2.2	Esquemas	20
2.3	Data Warehouse	22
2.3.1	Extrair, Transformar e Carregar	23
2.4	Mineração de dados	24
2.4.1	Descobrimento de Conhecimento em Banco de Dados	25
2.4.2	Descobrir conhecimento através da mineração	26
2.4.3	Algoritmos de mineração	27
3	CONSTRUINDO O <i>DATA WAREHOUSE</i>	30
3.1	Análise da Plataforma Nilo Peçanha	30
3.2	Criando o modelo dimensional	30
3.3	Arquitetura do <i>Data Warehouse</i>	32
3.4	Implementação do <i>Data Warehouse</i>	33
4	EXECUÇÃO DOS ALGORITMOS DE MINERAÇÃO	36
4.1	Bibliotecas utilizadas	37

4.2	Árvores de Decisão	37
4.3	Redes Neurais Artificiais	39
5	RESULTADOS E DISCUSSÕES	42
5.1	Árvores de Decisão	42
5.2	Redes Neurais Artificiais	45
6	CONCLUSÃO	47
	REFERÊNCIAS	49
	APÊNDICE A – ALGORITMO DE TRANSFORMAÇÃO PARA 2017	51
	APÊNDICE B – FLUXO DE TRANSFORMAÇÃO DOS DADOS DA PNP	53
	ANEXO A – DICIONÁRIO DE DADOS PNP 2017	54

1 INTRODUÇÃO

O ensino superior tem um papel importantíssimo na formação de profissionais capacitados, tendo os Institutos Federais e universidades como grandes agentes formadores, influenciando diretamente no desenvolvimento econômico e social do país. As inúmeras dificuldades enfrentadas por docentes e discentes nessas instituições somadas ao descaso governamental na falta de investimentos e sucateamento das instituições de ensino, cria um ambiente propício à evasão.

A Comissão Especial para Estudo da Evasão pela Secretaria de Educação Superior/Ministério da Educação e do Desporto (SESU/MEC ANDIFES, 1997) define que a evasão pode ser de três tipos: evasão do sistema, que acontece quando o aluno se ausenta definitivamente ou temporariamente do ensino superior; evasão da instituição, que se dá quando o aluno se desliga da instituição e a evasão de curso que é quando o aluno se ausenta do curso superior por diferentes motivos, sendo eles, desistência, abandono (não se matricular), exclusão por norma institucional, transferência ou reopção.

De acordo com Johann et al. (2012) é considerado evadido o aluno que, por qualquer motivo, rompa o vínculo jurídico estabelecido com a instituição de ensino, não renovando a sua matrícula no curso matriculado.

Filho et al. (2007) definem os conceitos de evasão anual média e evasão total, sendo a primeira calculada utilizando o total de alunos vinculados a um curso em um semestre para determinar quantos desses ainda não realizaram matrícula para o semestre subsequente. Já a evasão total é determinada pelo número de alunos que não concluíram sua formação em um número de anos, e é relacionada diretamente ao índice de titulação (porcentagem de formandos na totalidade de ingressantes do determinado ano).

A evasão gera inúmeras consequências negativas. Os alunos evadidos têm dificuldade para encontrar emprego e solidificar uma carreira sem uma qualificação. Além disso, afeta diretamente as taxas de rendimento escolar das instituições de ensino, diminuindo os números que representam a qualidade do ensino fornecido. É um assunto que vem sendo discutido desde o começo da expansão do ensino superior pelo mundo na década de 60. Desde então, diversos autores se dedicam a estudá-la, tentando entender quais causas eram essas e de que forma poderiam prevêê-las buscando uma maior permanência estudantil.

Este trabalho pretende identificar características relevantes na classificação de um perfil de aluno como propenso a evasão através da reorganização de dados educacionais e análise destes utilizando algoritmos de mineração.

1.1 MOTIVAÇÃO

O estudo dos números de evasão é essencial para tentar definir suas principais causas. Para isso, é necessária uma análise detalhada dos dados das instituições. Filho et al. (2007) falam como são raras as Instituições de Ensino Superior (IES) brasileiras que contam com algum programa institucional profissionalizado de combate à evasão. Os autores afirmam também a necessidade do estudo das suas causas, uma vez que não existe uma homogeneidade nos níveis de evasão para as diferentes áreas de ensino, impedindo a criação de um padrão a ser adotado.

Não só por falta de interesse um aluno desiste dos estudos, as causas da evasão são inúmeras e estão enraizadas na história do Brasil. De acordo com o relatório do UNICEF, Fundo das Nações Unidas para a Infância, a maioria dos alunos fora da escola, em todo o país, são pretos, pardos, indígenas e vivem em famílias com renda domiciliar per capita de até meio salário mínimo (UNICEF, 2021). Esses alunos muitas vezes abandonam os estudos para trabalhar, buscando ajudar financeiramente na situação difícil vivida em casa. Situações de trabalho infantil, gravidez na adolescência, repetidas reprovações e racismo estrutural também são grandes agentes no distanciamento dos alunos da escola.

Outro ponto crucial que alimenta a evasão é a falta de investimento, ano após ano o valor investido na Educação do país tem diminuído drasticamente. De acordo com o relatório de 2021 do INESC, Instituto de Estudos Socioeconômicos, o país atingiu o menor valor investido em educação desde 2012 e para 2022 o valor estimado será de R\$123,7 bilhões de reais, R\$ 6,2 bilhões menor que o valor para 2021 (INESC, 2021).

Alunos que antes tinham a oportunidade de estudar por receber algum auxílio ou docentes que tentam manter suas aulas interessantes lidando com desafios estruturais e financeiros, são diretamente prejudicados, criando uma série de problemas que impactam no crescimento do país como um todo. Quando não se há investimento, é extremamente difícil de se ofertar as melhores oportunidades de aprendizado, assim como criar um ambiente propício para que o aluno consiga desenvolver as competências necessárias para exercer sua cidadania, seu trabalho e sua vida pessoal.

Combater a evasão é um trabalho em conjunto: O estado deve garantir que os impostos sejam investidos em qualificação e remuneração dos profissionais, assim

como, em melhorias estruturais nas instituições, oferecendo um ambiente seguro e acolhedor que possibilite a permanência dos alunos na escola. Os docentes devem buscar didáticas que atendam as deficiências reais dos alunos, além de proporcionar a transformação desses em cidadãos. Resta também à família incentivar e garantir o bem estar e a assiduidade do aluno.

1.2 OBJETIVO GERAL

Motivado pelos importantes aspectos da evasão, esse trabalho tem como objetivo geral encontrar padrões e características da evasão através da mineração dos dados da PNP organizados em um *Data Warehouse*.

1.3 OBJETIVOS ESPECÍFICOS

- Organizar os dados da PNP em um *Data Warehouse* para facilitar a consulta e análise dos dados;
- Descobrir características relevantes na identificação de perfis com propensão a evasão através da mineração de dados utilizando árvores de decisão e redes neurais artificiais;
- Auxiliar a tomada de decisão de gestores em atividades contra a evasão com base nos resultados obtidos;

1.4 TRABALHOS RELACIONADOS

Considerando novamente que o tema da evasão escolar é um problema sério no nosso país, trabalhos relacionados com esse já foram desenvolvidos e chegaram a conclusões relevantes de acordo com seus objetivos específicos de pesquisa.

Em Souza et al. (2021) foi realizada uma abordagem supervisionada com o objetivo de prever precocemente quais estudantes teriam desempenhos insuficientes em uma disciplina realizada em modalidade de educação online, utilizando dados coletados do Google Sala de Aula e do Sistema Unificado de Administração Pública (SUAP), sistema acadêmico utilizado no IFPB, como fonte para a mineração de dados. Foram construídos modelos de classificação utilizando os métodos *Naive Bayes*, KNN, SVM, *Random Forest*, *Gradient Boosting* e *Extreme Gradient Boosting*.

Dutra et al. (2021) utiliza a Mineração de Dados Educacionais para prever estudantes com potencial de evasão, o trabalho também utilizou dados extraídos do

SUAP, que é o software de controle acadêmico e gestão utilizado no IFPB. A mineração foi executada através do Weka e utilizou os algoritmos *Part*, *OneR*, *J48* e *RandomTree*.

Esse trabalho se diferencia deles pela criação do *Data Warehouse* utilizando as informações da Plataforma Nilo Peçanha, além de realizar a mineração dos dados utilizando bibliotecas e algoritmos de classificação diferentes que serão explicados no capítulo 4.

1.5 METODOLOGIA

Neste subtópico serão descritas as atividades que foram desencadeadas com o desenvolvimento deste Trabalho de Conclusão de Curso (TCC). As atividades foram elaboradas tendo como base os livros Kimball e Ross (2013) e Kimball e Caserta (2004).

A primeira tarefa foi a análise da fonte original dos dados da Plataforma Nilo Peçanha, esses estão disponíveis em diferentes formatos. Um deles é a presente no site da PNP¹ que consiste em uma versão online dos dados separados por temas. Para este projeto foi escolhido o uso dos arquivos CSV uma vez que abrem um leque de opções de como trabalhar e analisar esses dados fugindo das limitações impostas pelo sistema de gráficos e tabelas presente na versão do site.

Na próxima atividade foram aplicados os conceitos de Modelagem Dimensional para criar uma modelagem que fizesse sentido para o nosso objetivo em questão, que é conseguir analisar esses dados com mais velocidade e eficiência.

Com o modelo dimensional estabelecido, o desenvolvimento da arquitetura do *Data Warehouse* foi iniciado, delimitando todas as características obrigatórias para que a esta funcionasse de forma completa e eficiente quando implementada.

Com o *Data Warehouse* desenvolvido, foi realizada a mineração dos dados utilizando algoritmos de classificação. Então os resultados da mineração foram analisados determinando as características mais importantes consideradas pelos algoritmos.

Por último este documento foi escrito evidenciando o desenvolvimento do *Data Warehouse*, os algoritmos e configurações utilizados e também levantando questões e afirmações sobre a evasão usando os resultados obtidos como base.

A seguir uma lista em ordem cronológica das atividades realizadas e uma

¹ plataformanilopecanha.mec.gov.br

descrição sobre o que foi realizado em cada uma delas:

- **Atividade 1 - Análise da Plataforma Nilo Peçanha:** Análise e definição de todas as características e limitações relevantes para a execução do trabalho;
- **Atividade 2 – Modelagem dimensional:** Elaboração de um modelo dimensional adequado que será utilizado para reorganizar os dados no *Data Warehouse*;
- **Atividade 3 – Data Warehouse:** Definição da arquitetura do *Data Warehouse* e desenvolvimento da solução de ETL usando o modelo dimensional já definido na atividade anterior;
- **Atividade 4 – Mineração dos dados:** Execução dos algoritmos de classificação (Árvores de Decisão e Redes Neurais) em busca da melhor configuração possível. Repetidas execuções das configurações escolhidas em busca da média dos resultados;
- **Atividade 5 – Conclusões:** Análise dos resultados da mineração de dados e elaboração de afirmações e questões relacionadas a evasão escolar;
- **Atividade 6 – Documento escrito:** Confecção do documento de Trabalho de Conclusão de Curso (TCC).

1.6 ORGANIZAÇÃO DO DOCUMENTO

No Capítulo 2 é explicado o referencial teórico, são abordados todos os conceitos relevantes para o escopo do trabalho, sendo eles a Plataforma Nilo Peçanha, Modelagem Dimensional, *Data Warehouse* e Mineração de Dados Educacionais utilizando algoritmos de classificação. No Capítulo 3 são evidenciados todos os processos de transformação dos dados e reorganização dos mesmos no modelo escolhido. Já no Capítulo 4 é explicada a metodologia de testes e mineração de dados utilizada. No Capítulo 5 são mostrados os resultados obtidos pelos algoritmos de classificação e quais informações podemos entender das métricas apresentadas. Por último temos o Capítulo 6 onde as conclusões do trabalho são apresentadas.

2 REFERENCIAL TEÓRICO

Este capítulo aborda os conceitos que foram utilizados como base para a produção deste trabalho. Primeiramente fala sobre o conceito de Modelagem Dimensional. Logo após, apresenta as definições de *Data Warehouse*, e por fim, fala sobre Mineração de Dados.

2.1 PLATAFORMA NILO PEÇANHA

A Plataforma Nilo Peçanha (PNP) surgiu com a proposta de ser “um ambiente virtual de coleta, validação e disseminação das estatísticas oficiais da Rede Federal de Educação Profissional, Científica e Tecnológica (Rede Federal)” (SETEC/MEC). Coletando todos os dados de matrículas, eficiência acadêmica, servidores e financeiros, durante o ano inteiro via Sistema Nacional de Informações da Educação Profissional e Tecnológica (Sistec). A plataforma atualmente possui dados referentes aos anos base de 2017 a 2021.

Iniciada em 2017, leva o nome do Nilo Procópio Peçanha que foi o primeiro Presidente da República negro e é considerado o patrono da educação profissional e tecnológica no Brasil. A plataforma tem como proposta organizar um conjunto de indicadores de desempenho, fortalecendo a gestão da educação pública tornando-a mais ágil e diminuindo a dificuldade em definir políticas públicas para aprimorar a Rede Federal.

Os indicadores produzidos servem para avaliar diversos aspectos da Rede Federal e também para incentivar as instituições de ensino a se apoderarem dessas informações, tornando possível o monitoramento das suas estatísticas e uma melhor visualização dos aprimoramentos necessários para atingir seus objetivos, desenvolver ferramentas e ações, assim como verificar os resultados destas. Outro ponto importante é a presença de diversos indicadores específicos que não estão presentes em outros censos e que sem eles seria bem mais difícil determinar se os objetivos determinados pelo governo estão sendo cumpridos.

No Guia de Referência Metodológica do MEC de 2021 (MEC et al., 2021), é explicado a importância da coleta e disponibilização destes dados para as instituições de ensino:

"Se não temos a contagem total de alunos dos Institutos Federais,

como determinar, por exemplo, se estão destinando 50% de suas matrículas para os cursos técnicos, 20% para os cursos de formação de professores (Lei 11.892/08) e 10% para os cursos do PRO-EJA (Decreto 5.840/06). Pelo mesmo motivo, torna-se impossível verificar se a Relação do número de Alunos por Professor, determinada pelo Plano Nacional de Educação (Lei 13.005/14), está sendo alcançada. Da mesma forma, sem o total de matrículas anuais, as instituições públicas não podem comunicar à sociedade qual tem sido o Gasto Corrente por Aluno (Acórdão TCU 2.267/05), dando consequência à diretriz de promover a transparência no uso do recurso público."(MEC et al., 2021, p. 6)

A plataforma contém um sistema de *Business Intelligence* (BI) onde é possível filtrar os dados de acordo com as opções disponíveis e analisar os dados exibidos através de tabelas e gráficos. Os dados da plataforma também estão disponíveis na plataforma governamental Dados Abertos¹ onde pode ser feito o download de arquivos CSV contendo todos os dados separados por ano e categoria.

Os microdados disponibilizados nos Dados Abertos são organizados e divididos em quatro categorias: matrículas, eficiência acadêmica, dados sobre os servidores e informações financeiras. A plataforma disponibiliza seus dados de duas formas: microdados em arquivos no formato CSV (*Comma-separated values* ou valores separados por vírgula em tradução livre) e *dashboards* com filtros e representações gráficas das informações coletadas a cada ano. Ambas as formas dificultam a realização de uma análise mais detalhada, uma vez que os arquivos CSV estão organizados de maneira diferente a cada ano e os *dashboards* entregam as informações de uma maneira pouco flexível, o que acaba impedindo a utilização de outras ferramentas para analisar os dados de formas mais específicas.

2.2 MODELAGEM DIMENSIONAL

De acordo com Kimball e Ross (2013), a Modelagem Dimensional é uma técnica amplamente utilizada para apresentar dados analíticos. É escolhida por fornecer, através de consultas rápidas, os dados de forma compreensível aos usuários, além de tornar os modelos dos bancos de dados mais simples, estruturados e limpos. Em Hokama et al. (2004) são citados motivos pelo qual os modelos dimensional se destacam em relação as outras alternativas.

"Como o modelo relacional trabalha com normalização, suas tabelas possuem menos registros e não têm redundâncias, apresentando assim uma melhor performance nas tarefas do dia a dia, como inclusões, alterações e exclusões de registros, mas ele só é adequado para consultas

¹ dadosabertos.mec.gov.br/pnp

simples de poucos registros. Para análises mais complexas com um universo de registros maior, o modelo dimensional oferece uma melhor alternativa, economizando em junções com várias tabelas, e armazenando dados que facilitam a análise das informações."(HOKAMA et al., 2004, p. 32)

A simplicidade é uma característica crítica na criação de um modelo dimensional, é ela que garante o entendimento dos dados, assim como, permite que o software entregue resultados rápidos eficientemente. O uso de modelos normalizados acaba escalando, criando modelos com ramificações complexas de difícil entendimento e memorização. A carga imprevisível das consultas dos usuários sobrecarrega o sistema pela quantidade de junções entre tabelas necessárias para conseguir uma informação, entregando uma péssima performance.

“Um modelo dimensional contém a mesma informação que um modelo normalizado, mas empacota os dados em um formato que entrega compreensibilidade, performance em consultas e resiliência a mudanças.” (KIMBALL; ROSS, 2013, p. 44, tradução nossa)

2.2.1 Elementos dos modelos dimensionais

Para se entender um modelo dimensional é necessário saber o que são alguns conceitos.

A tabela de fatos é uma tabela que possui medições numéricas associadas à visão de negócio e conexões para todas as dimensões relacionadas a essa visão. A sua chave primária é formada por meio da combinação das suas chaves estrangeiras de dimensão e nela deve-se tentar manter uma relação de um para um entre um evento de negócio determinado e uma única linha na tabela.

As tabelas de dimensão representam as perspectivas pelas quais as métricas são observadas. Kimball e Ross (2013) descrevem as tabelas de dimensão como o “quem, o que, onde, quando, como e porque” associados a um evento de negócio. Cada dimensão possui uma chave primária específica para garantir a integridade referencial com a tabela fato. As tabelas de dimensão são utilizadas para filtrar as consultas e montar as visões dos dados que são do interesse do negócio.

Um conceito também importante é o de grão, ele consiste em exatamente o que uma linha da tabela fato representa e é um aspecto importante do projeto e de difícil mudança no meio do desenvolvimento do projeto. Dessa forma deve ser bem pensado e definido, declarado antes da definição das dimensões e todas elas devem

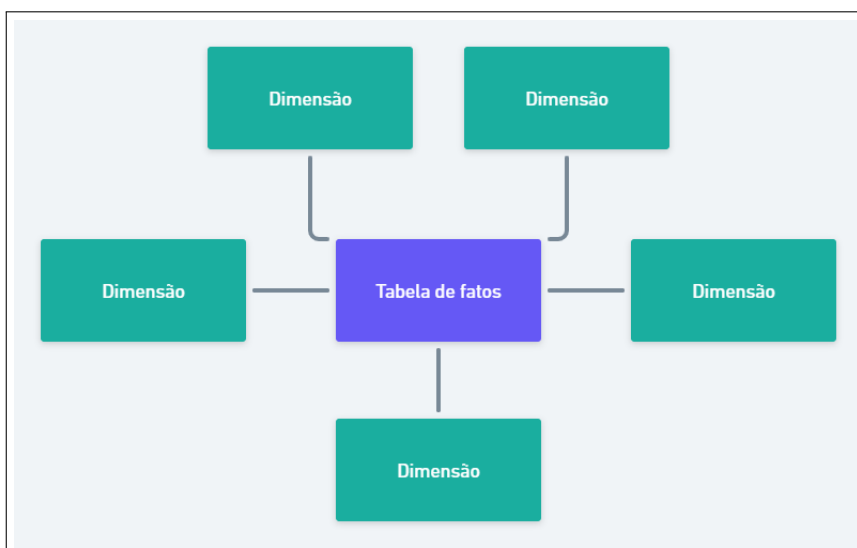
seguir a mesma granularidade. Um grão mal definido pode ser a ruína de um projeto de modelagem dimensional.

2.2.2 Esquemas

O esquema estrela possui uma modelagem simples de se abstrair e contém uma tabela de fatos central rodeada pelas tabelas de dimensão, justificando o seu nome. Possui uma hierarquia bem definida na qual a tabela de fatos central possui várias relações com outras tabelas enquanto essas possuem um único relacionamento com a tabela central. Através da sua simplicidade e simetria, usuários finais são diretamente beneficiados, uma vez que os dados são fáceis de entender e navegar, além dos benefícios de performance obtidos através de processos otimizados que de forma eficiente unem as tabelas com poucas junções. A sua arquitetura dimensional equivalente e simétrica torna o modelo mais previsível, o que ajuda na hora de realizar mudanças.

Na Figura 1 podemos ver um exemplo de esquema estrela, onde o retângulo roxo representa a tabela de fatos e os retângulos verdes representam os tabelas de dimensão.

Figura 1 – Exemplo de esquema estrela

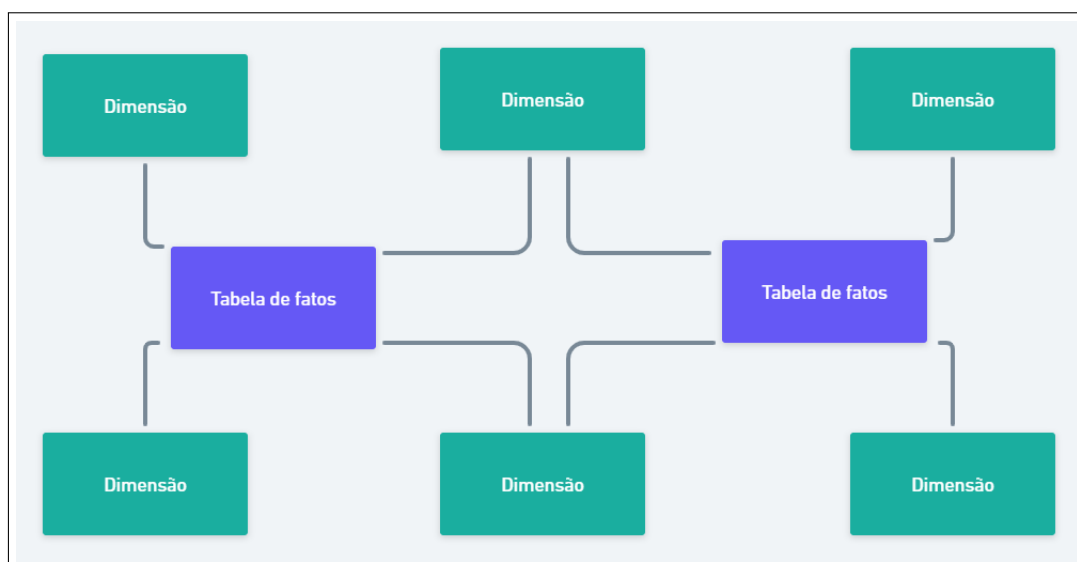


Fonte: Elaborado pelo autor

O esquema de constelação acontece quando o modelo possui múltiplos esquemas de estrela se relacionando. Esse possui várias tabelas fatos que compartilham

dimensões entre si, integrando diferentes assuntos dentro de um contexto de análise compartilhado. Na Figura 2 é representado um exemplo de esquema constelação, onde os retângulos roxos representam as tabelas de fatos e os retângulos verdes representam as tabelas de dimensão.

Figura 2 – Exemplo de esquema constelação

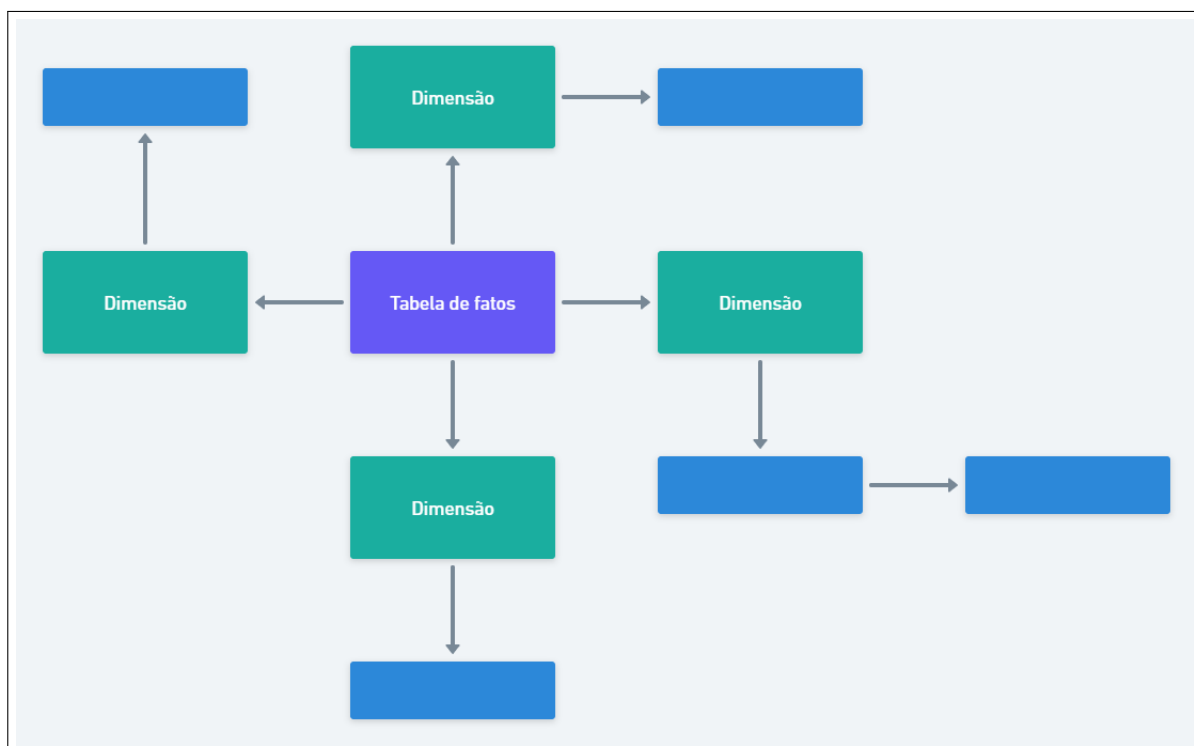


Fonte: Elaborado pelo autor

Já o esquema de floco de neve representa uma extensão do esquema de estrela e suas dimensões possuem ramificações que possibilitam a remoção das repetições existentes nos dados. Uma hierarquia se desenvolve, aumentando a quantidade de relacionamentos entre as tabelas, consequentemente aumentando a complexidade do modelo. Não traz muitas vantagens e o ganho de espaço não é relevante.

A Figura 3 representa um exemplo de esquema floco de neve onde o retângulo roxo representa a tabela de fatos, os retângulos verdes representam as tabelas de dimensão e os retângulos azuis representam as ramificações das dimensões.

Figura 3 – Exemplo de esquema floco de neve



Fonte: Elaborado pelo autor

Para a realização desse trabalho foi escolhido o esquema estrela por possuir a organização mais coerente com os dados selecionados e com o objetivo do *Data Warehouse* desenvolvido.

Ainda de acordo com Kimball e Ross (2013), um projeto de modelagem dimensional deve seguir os seguintes passos para obter sucesso: Primeiramente deve-se determinar o processo de negócio e quais suas características e especificidades. Após isso deve-se declarar a granularidade, definindo o que será representado por cada linha da tabela de fatos. Por último pode-se identificar as dimensões, reorganizando as informações ao redor da tabela de fatos. O modelo dimensional desenvolvido seguindo estes passos será apresentado e explicado no Capítulo 3.

2.3 DATA WAREHOUSE

Data Warehouse (DW) ou Armazém de Dados, em tradução livre, é um sistema de gerenciamento de informações com objetivo de adicionar valor a dados ao transformá-los em ativos que podem ser usados diretamente para influenciar decisões.

Ele é conceituado pelo sistema ETL, que significa *Extract-Transform-Load* (Extrair-Transformar-Carregar em tradução livre) ou ECCD que significa *Extract-Clean-Conform-Deliver* (Extrair-Limpar-Adequar-Entregar em tradução livre), ambos norteiam a arquitetura de como os dados são extraídos, limpos, armazenados e entregues de forma simples através de consultas. (KIMBALL; ROSS, 2013)

Diferente dos sistemas que usam bases de dados normalizadas que possuem inúmeras transações constantes, principalmente inserções, atualizações e remoções, gerando um fluxo de atualização onde poucos registros são modificados por vez, um ambiente informacional, como um *Data Warehouse*, foca em transações de leitura, gerando um número de requisições menores, mais longas e mais complexas.

A integração também é uma característica muito importante do *Data Warehouse*, é a etapa onde acontece a padronização dos dados retirados da fonte original, tendo grande impacto na manutenção da singularidade das informações assim como na remoção de inconsistências que possam existir nos dados.

Para construir um bom *Data Warehouse*, é necessária uma análise de requisitos completa que defina todas as realidades e limitadores dos dados. Também é necessário definir qual será a utilidade e como eles devem ser organizados antes de serem entregues à aplicação final. Após isso é possível definir uma arquitetura coesa que seja capaz de realizar todos os passos do sistema de forma eficiente.

2.3.1 Extrair, Transformar e Carregar

Em Kimball e Caserta (2004) é afirmada a importância do processo de ETL, sendo responsável por remover erros, corrigir dados incompletos, promover métricas para o nível de confiança dos dados, capturar o fluxo de transações, ajustar múltiplas fontes para serem usadas juntas e estruturar os dados para se tornarem usáveis por outras ferramentas. Como é dito em seu nome, o processo vai extrair, limpar, normalizar e entregar os dados utilizando uma série de operações.

O primeiro passo para o processo de transformação é extrair os dados da fonte original. Informações provenientes de diversas fontes e que estão armazenadas e organizadas de maneiras totalmente diferentes podem ser extraídas e reorganizadas em um *Data Warehouse* utilizando a tecnologia de comunicação adequada para o respectivo tipo de sistema.

Após isso, as informações foram transferidas para a *staging area*, ou área de preparação, onde sofreram transformações para torna-las uniformes e verificar se são válidas. A utilização da área de preparação traz vantagens como a recuperabilidade e

auditoria dos dados. É um espaço proibido aos usuários, uma vez que a interferência pode causar inconsistências. Em pequenos projetos de *Data Warehouse*, como este, tabelas independentes são o suficiente para montar a área de preparação.

Depois de preparados, os dados são transformados de acordo com a modelagem dimensional especificada para o projeto onde esta será a parte do *Data Warehouse* que contém os esquemas estrela, cubos analíticos, *data marts*, dentre outros, buscando ser o mais próximo possível do negócio. Essas escolhas impactam diretamente na forma como os dados serão guardados, recuperados, ordenados, filtrados, atualizados e agregados. O *Data Warehouse* possui diversas operações direcionadas aos dados, como sumarizações, agregações e análises multidimensionais.

A agregação é o processo de agrupamento de dados similares. De forma geral, agrupa todos os registros relacionados, depois disso, os agrupa em estruturas próprias e os agrega de acordo com a lógica do problema, buscando responder à questão levantada, obter vantagens e crescimento.

Diversos formatos de arquivos podem ser escolhidos como formato final após as transformações, sendo eles *flat files*, *XML*, *JSON*, *Datasets*, *XML Schema*, tabelas relacionais ou tabelas independentes. A alta flexibilidade da tecnologia em se adaptar aos mais diferentes projetos, trazendo bons resultados, além também da capacidade de manipular quantidades enormes de dados, são motivos que justificam o crescimento explosivo no uso dela.

2.4 MINERAÇÃO DE DADOS

Witten et al. (2011) fala sobre o que é e como acontece a mineração de dados:

“Mineração de dados é a técnica de resolver problemas através da análise de bancos de dados já existentes, identificando padrões e informações presentes nos modelos criados através de análises que são de maneira geral automatizadas por um computador utilizando os dados que estão armazenados digitalmente.” (WITTEN et al., 2011, p. 4, tradução nossa)

Mais do que nunca os sistemas de informação guardam uma quantidade exorbitante de dados, dados estes muitas vezes que estão desorganizados e que por isso acabam por não terem serventia. Com a ajuda de conceitos e tecnologias que já falamos anteriormente, utilizando a Modelagem Dimensional e a arquitetura do *Data Warehouse*, é possível criar um ambiente perfeito com dados organizados de acordo

com um objetivo para que possam ser analisados. Como é abordado em (NINGNING, 2010):

“Combinar *Data Warehouse* com mineração de dados organicamente vai melhorar grandiosamente as capacidades de organização e utilização das informações tornando-as melhores para tomar decisões. O *Data Warehouse* cria melhores fundações para a mineração de dados, incluindo extração, limpeza, ordenação e processos de consistência” (NINGNING, 2010, p. 437, tradução nossa)

2.4.1 Descobrimento de Conhecimento em Banco de Dados

O *Knowledge Discovery in Databases* (KDD), ou Descobrimento de Conhecimento em Bancos de Dados, consiste em uma série de requisitos que devem ser seguidos para criar um ambiente propício para o descobrimento de padrões e tendências dos dados escolhidos para análise.

Dentre os requisitos estão os de identificação da perspectiva do cliente, ou seja, como o projeto funciona e quais seus objetivos, o entendimento dos domínios da aplicação e qual o conhecimento requisitado, processos executados de forma detalhada quando se é feita a Modelagem Dimensional dos dados. Também é necessário a seleção, limpeza, processamento e remodelação dos dados de fontes externas de acordo com a modelagem, atividades executadas pelo *Data Warehouse*.

“Mineração de dados é sobre resolver problemas analisando informações já presente em uma base de dados. Suponha que o problema é a inconstância na fidelidade dos clientes em um mercado altamente competitivo. Uma base de dados de escolhas dos clientes unida de perfis destes é a solução para esse problema.” (WITTEN et al., 2011, p. 4, tradução nossa)

Witten et al. (2011) também afirma que a mineração de dados está presente em diversos tipos de aplicações e com escopos totalmente diferentes. Ela pode ser usada para minerar a internet por informações, ajudar em decisões envolvendo julgamento, criação de imagens espaciais, pré-carregamento de informações, diagnósticos médicos, marketing e vendas, além das inúmeras aplicações científicas.

Desta forma, a criação de um *Data Warehouse* com as informações de matrículas de alunos da rede federal e suas situações será de grande utilidade para encontrar e descrever padrões presentes entre os alunos evadidos.

2.4.2 Descobrir conhecimento através da mineração

Para realizar as análises destes dados, além de organizá-los, é necessário a utilização de algoritmos de Inteligência Artificial e *Machine Learning*, estes serão escolhidos com base nas necessidades e objetivos do projeto e ficarão responsáveis por realizá-las, uma vez que é humanamente impossível processar dados desta dimensão por conta própria. Além disso, possuem a capacidade de melhorar os processos trazendo mais eficiência e diminuição nos custos, assim como mais competitividade, produtividade e flexibilidade para lidar com as mudanças no decorrer do projeto. Ningning (2010) afirma sobre a mineração de dados:

“Pode ser dito que mineração de dados é o processo de construção de modelos e descoberta de relações dentre dados em grandes dimensões utilizando diversos métodos e ferramentas analíticas. Estes modelos e relações podem ser utilizados para decisões e predições”. (NINGNING, 2010, p. 437, tradução nossa)

A mineração de dados pode ser utilizada para uma grande quantidade de tipos de análises além do aprendizado de máquina e inteligência artificial, sendo elas comparativas, de tendências ou estatísticas, sendo capazes de gerar inúmeras agregações e fatos diferentes que formam, solidificam e apoiam visões do negócio.

Os processos de mineração e análise dos dados também podem ser aplicados para o desenvolvimento de pesquisas científicas, muitas vezes utilizados para analisar dados de atletas de alta performance, na análise de dados de administrações públicas, processamento de dados médicos para previsão de doenças ou em qualquer segmento que seja possível o melhor conhecimento e evolução a partir dessas análises.

Assim como na extração de minerais, os processos de mineração de dados também têm o objetivo de extrair conhecimentos ricos em valor dentre uma imensidão de dados, a exploração dessa gama de informações em perspectivas complexas de comparação e previsão são capazes de transformar informações antes sem valor em verdadeiras pedras preciosas.

Mesmo com todas as automações e inteligências artificiais, o processo de mineração de dados deve caminhar de mãos dadas com a concepção humana, uma vez que por mais avançadas que sejam, as tecnologias ainda não possuem certos aspectos de personalidade, intuição e subjetividade em suas decisões, só assim o conhecimento gerado terá uma maior credibilidade, respaldando decisões com menos risco, além da detecção e solução de problemas com maior agilidade.

2.4.3 Algoritmos de mineração

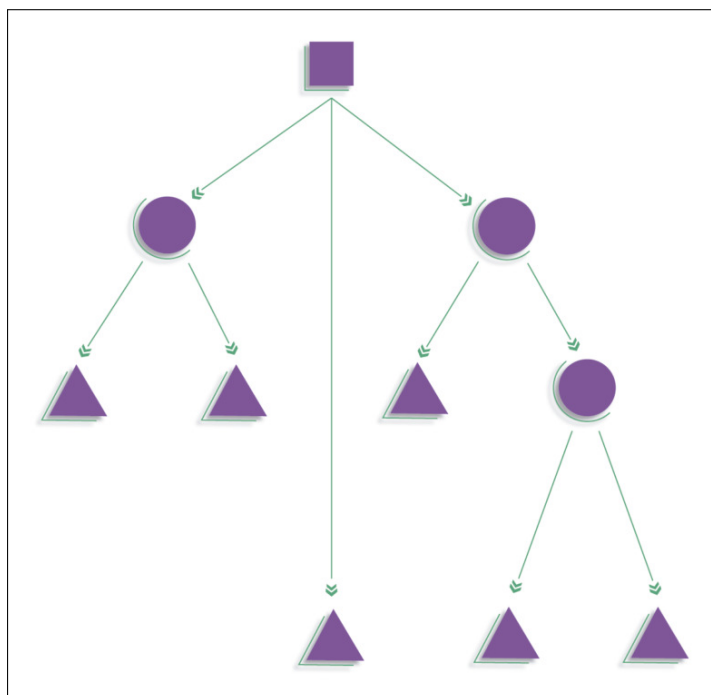
Os algoritmos de mineração podem ser divididos em três categorias, os supervisionados, os não supervisionados e os de aprendizagem por reforço. Cada um dos tipos se adéqua a diferentes necessidades.

Nesse trabalho foram utilizados os algoritmos supervisionados, esses podem ser subdivididos em algoritmos de classificação e algoritmos de regressão. Os de classificação tem como propósito classificar amostras levando em consideração os dados informados. Já os de regressão trabalham com a previsão de valores através da relação entre as características fornecidas.

A árvore de decisão é um algoritmo de aprendizado de máquina, é amplamente utilizado na área da Mineração de Dados por possuir um grande potencial de melhora na tomada de decisões levando em consideração os dados mais relevantes para o contexto de um problema. O modelo da árvore de decisão representa e organiza os dados de uma maneira hierárquica, onde os dados de maior relevância são colocados no primeiro nó e o restante é distribuído nos nós seguintes, também chamados de folhas. Isso cria uma visualização objetiva que facilita sua compreensão. Em Devi e Nirmala (2013) são explicados os elementos que fazem parte dessa estrutura.

“Árvores de decisão são poderosas e populares ferramentas para classificação e predição. Elas representam regras nas quais podem ser entendidas por humanos e usadas em sistemas de conhecimento como bases de dados. [...] O objetivo é criar um modelo que prediz o valor de uma variável alvo baseada em inúmeras variáveis de entrada. Uma árvore de decisão é um fluxograma na estrutura de uma árvore, onde cada nó interno denota um teste em um atributo, cada galho representa o resultado do teste e cada folha (últimos nós) contém um rótulo de classe. O nó mais alto é chamado de raiz. Uma árvore pode ser “aprendida” através da divisão da fonte em menores conjuntos baseados num teste de atributo. O processo é repetido em cada subconjunto derivado de forma recursiva chamando o particionamento recursivo. A recursão é completada quando o subconjunto de um nó possui o mesmo valor da variável alvo, ou quando a divisão não mais adiciona valor para as predições. Na mineração de dados, árvores de decisão podem ser descritas como a combinação de técnicas matemáticas e computacionais para ajudar na descrição, categorização e generalização de um conjunto de dados.” (DEVI; NIRMALA, 2013, p. 343, tradução nossa)

Na Figura 4 é mostrada uma representação de uma árvore de decisão onde o quadrado representa o nó raiz, os círculos representam os nós folhas, as setas representam as ramificações e o triângulo representa o resultado final obtido.

Figura 4 – Nós e Ramificações das Árvores Decisórias

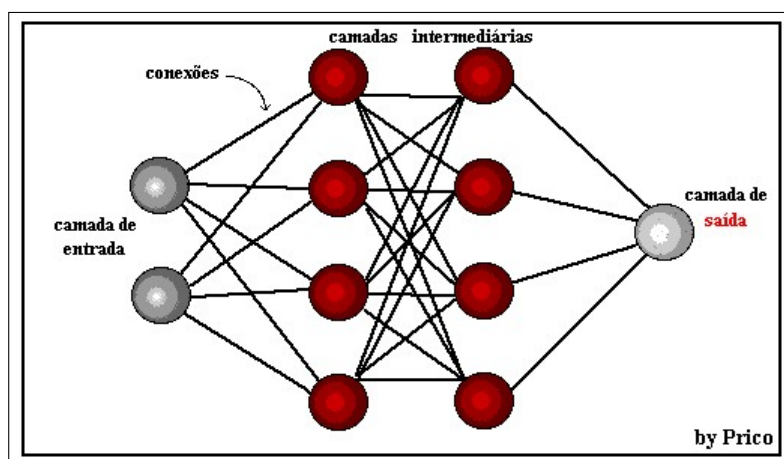
Fonte: (NAPOLEÃO, 2018)

Rede Neural Artificial é um algoritmo de aprendizado de máquina e recebe esse nome por sua estrutura ser inspirada no cérebro humano, tentando replicar a forma como os neurônios compartilham sinais entre si. Nesse trabalho implementamos redes neurais com perceptron multicamadas, algoritmo que treina utilizando retropropagação e que de forma progressiva utiliza vários neurônios artificiais e várias camadas que se comunicam até que um resultado seja obtido. Como é explicado em Team (2020):

“Um único neurônio artificial recebe um sinal e depois transfere-o para os outros neurônios ocultos após processamento e treino. [...] A retropropagação ajuda na formação de redes neurais artificiais. Quando as redes neuronais artificiais se formam, os valores dos pesos são atribuídos aleatoriamente. O utilizador estabelece pesos aleatórios porque não está ciente dos valores correctos. Quando o valor é diferente da rede de alimentação esperada, considere-o como um erro. O algoritmo é definido para que o modelo altere os parâmetros cada vez que a saída não é a esperada. O erro tem uma relação com redes neurais artificiais. Assim, quando o parâmetro muda, o erro também muda até a rede neural encontrar a saída desejada através do cálculo do gradiente de descida.” (TEAM, 2020)

Na Figura 5 é possível observar a organização em camadas construída por uma rede neural artificial. Na camada de entrada temos os dados e padrões apresentados à rede. Nas camadas intermediárias, também conhecidas como *hidden layers*, ou camadas escondidas em tradução livre, é onde acontece a maior parte do processamento através de suas conexões, tem como objetivo a extração de características. Na camada de saída é onde o resultado final é concluído e apresentado (CARVALHO, 2009).

Figura 5 – Organização em camadas



Fonte: (CARVALHO, 2009)

3 CONSTRUINDO O *DATA WAREHOUSE*

Nesse capítulo são abordadas as atividades realizadas para desenvolver o *Data Warehouse*, desde a análise inicial dos dados da PNP até a implementação do modelo dimensional no banco de dados *PostgreSQL*.

3.1 ANÁLISE DA PLATAFORMA NILO PEÇANHA

Ao analisar as opções de formato que os dados da PNP estão disponíveis, foi definido que os arquivos CSV seriam utilizados uma vez que abrangiam diferentes possibilidades para utilização dos dados.

Também foi realizada uma pesquisa para verificar a existência de alguma API pública que já trabalhasse os dados da plataforma, onde nada foi encontrado. Trabalhos relacionados encontrados utilizaram *softwares* específicos para analisar os arquivos CSV diretamente.

Para que fosse possível analisar uma quantidade maior de dados, foram selecionadas as informações que estavam presentes em todas as edições lançadas pela PNP. Os dados escolhidos pertencem à categoria "Matriculas". O Anexo A apresenta o dicionário de dados da PNP referente ao ano de 2017. Os dados escolhidos para compor o *Data Warehouse* serão descritos posteriormente.

Como um dos objetivos deste trabalho é facilitar e agilizar o acesso aos dados da plataforma, um modelo dimensional se torna necessário.

3.2 CRIANDO O MODELO DIMENSIONAL

Seguindo os passos recomendados em Kimball e Ross (2013), foi determinado o processo de negócio que é o controle da situação dos alunos. Também foi definida a granularidade do projeto, sendo ele uma unidade de matrícula. Foram identificadas as dimensões abordando as perspectivas de curso, perfil do aluno e ciclo. E também foi definido que a tabela fato seria a tabela matrícula.

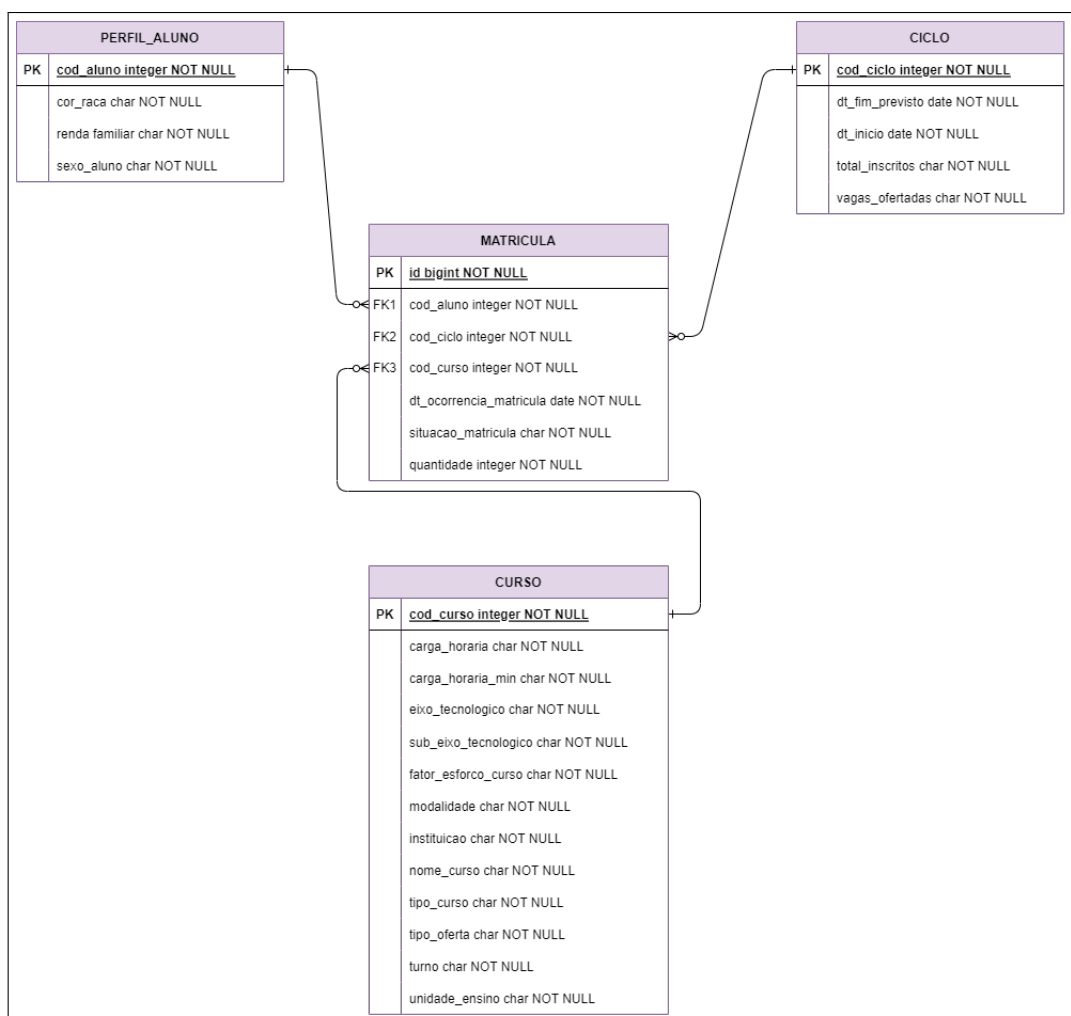
Considerando as características específicas de cada esquema, foi escolhido que o modelo dimensional do trabalho seguiria o Esquema Estrela. Sendo este o que mais se encaixou no escopo do desenvolvimento do trabalho, uma vez que criou uma arquitetura simples e de rápido acesso, o que resultou em alta performance

na realização de consultas mais complexas, características também explicadas em Hokama et al. (2004).

“O esquema estrela é uma estrutura simples, com poucas tabelas e ligações (relacionamentos) bem definidas [POE, KLAUER, BROBST, 1998], assemelha-se ao modelo de negócio, o que facilita a leitura e entendimento, não só pelos analistas, como por usuários finais não familiarizados com estruturas de banco de dados. Permite a criação de um banco de dados que facilita a execução de consultas complexas, podendo ser realizadas de modo eficiente e intuitivo pelo usuário.” (HOKAMA et al., 2004, p. 32)

Na Figura 6 é exibida a modelagem dimensional do esquema estrela e esta representa como as tabelas estão organizadas e se conectam dentro do banco de dados.

Figura 6 – Modelo dimensional - Esquema estrela



Fonte: Elaborado pelo autor

A tabela central MATRÍCULA representa a tabela de fatos, tabela esta que agrupará todas as chaves estrangeiras das tabelas de dimensão. Além disso, ela guardará a medição de quantidade calculadas em relação direta com seu grão.

A tabela de dimensão CURSO possui todas as informações referentes aos cursos e suas instituições, como carga horária, eixo técnico, modalidade, tipo do curso, turno e etc. A tabela PERFIL_ALUNO agrupa as informações pessoais referentes a todos os perfis de alunos, sendo o agrupamento das informações de raça, renda familiar e sexo. Já a tabela CICLO guardará as informações referentes aos ciclos de matrículas, contendo datas de início, fim previsto, totais de inscritos e vagas.

3.3 ARQUITETURA DO *DATA WAREHOUSE*

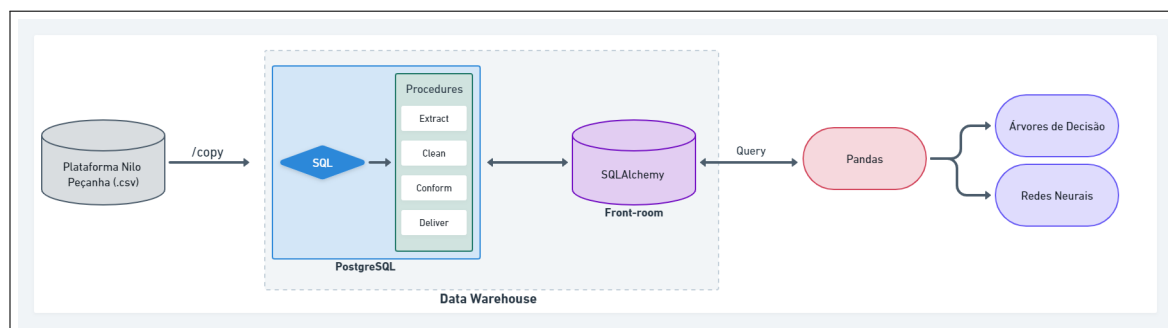
O primeiro passo para o desenvolvimento de um Data Warehouse, de acordo com Kimball Kimball e Caserta (2004), é analisar a sua fonte original de dados, ver quais são suas limitações e determinar como a comunicação será estabelecida com o armazém. Os dados da Plataforma Nilo Peçanha estão distribuídos em arquivos CSV, que são arquivos que possuem seus valores separados por vírgula. Este formato não é amigável para realizar análises de aprendizado de máquina ou inteligência artificial. Além de que, há divergências na forma como os campos são nomeados em edições de outros anos.

Após analisar e determinar as modificações necessárias, os dados foram enviados para uma área de preparação onde serão criadas tabelas independentes, uma para cada arquivo CSV. Elas possuirão os dados originais e servirão como backup em caso de necessidade de recarregar os dados, evitando uma releitura dos arquivos CSV. Estas tabelas independentes serão utilizadas no processo de ETL (*Extract-Transform-Load*) onde os dados serão transformados e reorganizados de acordo com o modelo dimensional, o esquema estrela já definido, e poderão ser carregados através de uma interface de comunicação com o *Data Warehouse*.

Na Figura 7 está representada a arquitetura que foi definida para o trabalho. O cilindro cinza representa a base de dados obtida da PNP. O quadrado cinza representa o *Data Warehouse*, dentro dele o quadrado azul possui elementos que representam as ações executadas utilizando *procedures* SQL. O cilindro rosa representa a biblioteca utilizada para se comunicar com os algoritmos de mineração através de consultas SQL. O círculo vermelho representa a biblioteca utilizada para transformar os dados no

formato requerido pelos algoritmos de mineração, representados pelos círculos roxos.

Figura 7 – Arquitetura do trabalho



Fonte: Elaborado pelo autor

A utilização da biblioteca SQLAlchemy com o *Data Warehouse* tem como objetivo facilitar a execução de consultas aos dados, possibilitando uma comunicação mais rápida e eficiente para os processos de análise.

3.4 IMPLEMENTAÇÃO DO *DATA WAREHOUSE*

Após definir a arquitetura do *Data Warehouse* e a Modelagem Dimensional que foi seguida para reorganizar os dados da PNP, o próximo passo é iniciar a manipulação dos dados presentes na fonte.

Foi escolhido o banco de dados *PostgreSQL* uma vez que além de ter seu código aberto e apresentar ótimos resultados em relação a performance de inserção e consulta de dados, possui funcionalidades de registro de *logs* antecipados que o torna mais tolerante a falhas, assegurando com mais facilidade a integridade dos dados (ARORA, 2022).

Primeiramente, todos os arquivos CSV foram convertidos para o padrão UTF-8, uma vez que da forma que estavam disponibilizados, a função necessária para a cópia dos dados não funcionaria. Dito isso, foi utilizado a função *COPY* do PostgreSQL para transferir rapidamente os dados dos CSV para suas respectivas tabelas, criadas para cada um dos anos. Foram utilizadas quatro tabelas diferentes pois as informações estão organizadas em posições diferentes, com diferente quantidade de campos nos arquivos CSV de cada ano. A função de cópia necessita que as informações estejam na mesma sequência e quantidade da fonte para que seja executada com sucesso.

Para a execução desse passo, o *container* que contém a instância do banco de dados foi acessado via linha de comando, então o cliente Postgres foi iniciado e a função *COPY* foi executada para cada uma das tabelas já criadas para cada ano. No Algoritmo 1 é exibido um pseudocódigo da função utilizada, onde na linha três é definido o símbolo que será usado para delimitar os campos do arquivo lido.

Algoritmo 1 – Pseudocódigo da função COPY

```

1 COPY NOME_DA_TABELA (CAMPO1, CAMPO2, CAMPO3, ...)
2 FROM './caminho_do_arquivo/arquivo.csv'
3 DELIMITER ';' CSV HEADER;
```

Fonte: Elaborado pelo autor

As tabelas de dimensão foram criadas fazendo o agrupamento das informações originais por seus campos já estabelecidos na modelagem. Dessa forma conseguimos criar entidades específicas de perfil de aluno, curso e ciclo. Gerando uma grande redução na quantidade de dados necessários para representar a mesma informação.

Essa transformação foi realizada com a criação de um procedimento em SQL (*Structured Query Language*, ou Linguagem de Consulta Estruturada), seguindo um fluxo para agrupar e inserir os dados de cada um dos anos em uma única tabela, agrupando essas informações pelos campos escolhidos que estavam disponíveis em todas as edições dos microdados. Após a execução para cada um dos anos, foi realizando o reagrupamento dessa tabela para que qualquer registro repetido fosse removido, atualizando o campo quantidade para os outros registros relacionados.

Nesse primeiro momento também foram utilizadas funções para normalização e limpeza dos dados. Utilizando funções do PostgreSQL como *UNACCENT()* para remover acentos das palavras, *UPPER()* para deixar todas as letras em maiúsculo e *REPLACE()* para substituir símbolos como pontos e vírgulas em dados numéricos, foi possível gerar uma base de dados mais limpa e coesa, possibilitando que os reagrupamentos fossem eficazes. Um exemplo do código utilizado na transformação pode ser observado no Apêndice A.

Dando sequência ao procedimento, os dados reagrupados na tabela final agora estão prontos para serem divididos em suas respectivas dimensões. Essa parte se deu pelo preenchimento das tabelas com o agrupamento da tabela final por cada um dos campos das dimensões. Cada tabela de dimensão possui um código único como chave primária para possibilitar a conexão com a tabela central de fatos.

Por fim, a tabela de matrículas foi preenchida pesquisando em cada dimensão qual o código do detalhamento que representa aquele perfil de matrícula, populando

a tabela fato com essas chaves estrangeiras e as informações de data de ocorrência da matrícula, situação desta e a quantidade de matrículas que se encaixam nessa descrição.

No Apêndice B estão descritas as tabelas criadas para esse processo com seus respectivos campos, assim como uma representação do fluxo executado.

Em comparação os quatro arquivos CSV somados possuem mais de 4 milhões de linhas, um volume exorbitante para a quantidade de consultas que serão realizadas. Dessa forma após a transformação no modelo estrela as tabelas passaram a ter os seguintes tamanhos: A dimensão de perfis de aluno conta com 123 registros, a dimensão de curso após transformada conta com 39.624 registros, a de ciclo se tornou 30.320 registros e a de MATRICULA conta com 1.458.217 registros, conectando todas as dimensões e calculando a quantidade de matrículas que se encaixam em cada linha.

Com a redução no tamanho das tabelas se torna mais fácil e rápido de realizar consultas específicas diminuindo a quantidade de conexões entre as tabelas necessárias para chegar ao resultado desejado. Característica importante para o tipo de análise que pretendemos realizar.

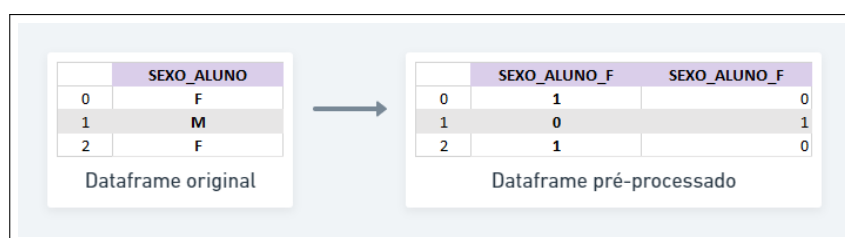
4 EXECUÇÃO DOS ALGORITMOS DE MINERAÇÃO

Nesse capítulo iremos abordar os detalhes na execução dos algoritmos de mineração, primeiramente falaremos sobre as bibliotecas utilizadas e depois detalharemos a execução dos algoritmos de Árvores de Decisão e Redes Neurais Artificiais.

Para executar os algoritmos de mineração foram utilizados os dados das matrículas de cursos superiores com data de ocorrência do ano de 2017. Sendo assim, considerando também apenas os registros com o tipo de curso bacharelado, especialização lato sensu, licenciatura e tecnólogos. Os registros com situação de matrícula igual a abandono, cancelada, reprovada, transferência externa e transferência interna foram considerados como evadidos. Já para os registros não evadidos, foram considerados apenas os com situação de matrícula iguais a concluída.

Antes de iniciar os algoritmos é feito o pré-processamento do resultado da consulta ao *Data Warehouse*, este é transformado em um formato adequado para a análise do algoritmo. Utilizando a função *get_dummies* da biblioteca *pandas*¹, cada valor diferente de cada coluna do *dataframe* é transformado em uma nova coluna onde receberá o valor de 0 ou 1, sendo 0 quando aquele registro não possuir a característica e 1 quando possuir. Um exemplo dessa transformação pode ser observada na Figura 8, onde um *dataframe* com as informações de sexo do aluno é transformado.

Figura 8 – Pré-processamento do *dataframe*



Fonte: Elaborado pelo autor

Para cada iteração do laço de repetição, é utilizada a função *train_test_split* para executar a divisão do *dataframe* retornado pela *query* com os campos selecionados para análise. Essa função recebe como parâmetros os valores "X" que representam

¹ <https://pandas.pydata.org>

os campos a serem considerados na formação da árvore, os valores “Y” que são os valores alvos da predição e um *test_size*, representando a porcentagem do conjunto de dados que deve ser usado para treinar o modelo.

4.1 BIBLIOTECAS UTILIZADAS

O *Scikit-learn*² é uma biblioteca de Aprendizado de Máquina desenvolvida em Python. Ela possui ferramentas simples e eficientes para análises de dados preditivas. Foi escolhida por ser gratuita, reutilizável em diversos contextos e construída utilizando famosas bibliotecas como *NumPy*³, *SciPy*⁴ e *matplotlib*⁵.

Se mostrou eficiente e sucinto ao oferecer com simplicidade quase todas as ferramentas necessárias para a execução das análises propostas por esse trabalho. Dentre a sua vasta quantidade de opções, escolhemos trabalhar com algoritmos de classificação utilizando as Árvores de Decisão e as Redes Neurais Artificiais, uma vez que buscamos identificar alunos que possam potencialmente evadir e estes atendem essa proposta.

Foi utilizada a biblioteca *SQLAlchemy*⁶ para criar uma *engine* de conexão com o nosso *Data Warehouse*, possibilitando que a biblioteca *pandas* fosse usada em conjunto para executar *queries* de consulta com retornos no formato de *dataframes*, formato este que facilita a organização dos dados nas diferentes formas requeridas pelos algoritmos de mineração. Já a biblioteca *matplotlib* foi utilizada para gerar as imagens utilizadas para representar os resultados obtidos.

4.2 ÁRVORES DE DECISÃO

Para a implementação das árvores de decisão com o *Scikit-learn*, foi utilizado a classe *DecisionTreeClassifier* que possui uma série de parâmetros configuráveis. São necessários testes com diferentes parâmetros em busca de refinar os resultados obtidos. Para escolher as configurações que mais beneficiam o nosso conjunto de dados, foi criado um laço de repetição contendo as opções para os parâmetros escolhidos da classe disponibilizada pela biblioteca.

Todos os parâmetros escolhidos e seus respectivos valores possíveis podem ser observados no Quadro 1.

² <https://scikit-learn.org>

³ <https://numpy.org>

⁴ <https://scipy.org>

⁵ <https://matplotlib.org>

⁶ <https://sqlalchemy.org>

Quadro 1 – Parâmetros escolhidos e seus possíveis valores para Árvore de Decisão

Parâmetro	Valores
test_size	[0.1, 0.2, 0.3, 0.4, 0.5]
max_depth	[2, 3, 5, 8, 13, 21, 34]
min_samples_splits	[0.1, 0.2, 0.3, 0.4, 0.5]
min_samples_leafs	[0.1, 0.2, 0.3, 0.4, 0.5]
random_state	[None, 1]
criterion	['gini', 'entropy']
splitter	['random', 'best']

O parâmetro *max_depth* define a profundidade máxima que a árvore de decisão pode chegar. Os parâmetros *min_samples_splits* e *min_samples_leafs* determinam a porcentagem de registros mínimos que deve existir para que a divisão de um nó seja feita e qual a quantidade mínima de registros nas folhas da árvore, respectivamente. O *random_state* controla a randomicidade do estimulador, quando ativado sempre permutará as variáveis de entrada randomicamente. O *criterion* define a função de medida de qualidade das divisões, utilizando a medida de impureza Gini quando igual a “gini” e utilizando ganho de informação quando igual a “entropy”. Já o *splitter* determina a estratégia usada para realizar a divisão em cada nó, podendo ser “best” para encontrar a melhor divisão e “random” para escolher a melhor divisão randômica.

Os parâmetros *max_depth*, *min_samples_splits* e *min_samples_leafs* têm uma grande importância, uma vez que quando bem configurados evitam que aconteça o *overfitting* da árvore de decisão, que é quando um modelo se adapta muito bem na parte do treino, mas não performa com a mesma qualidade quando é feito o teste com novos dados.

No Algoritmo 2 está representado o pseudocódigo utilizado para executar o laço de repetição descrito. Depois que cada árvore de decisão é criada, treinada e realizada a predição, um registro com suas informações de parâmetros e resultados é adicionado a uma tabela do banco de dados. Nessa tabela foram guardadas as informações relevantes para a escolha de qual árvore de decisão funciona melhor com o nosso conjunto de dados, sendo estas: Rótulos das variáveis de entrada, importância de cada uma delas dentro do modelo, acurácia, precisão, *recall*, profundidade e número de folhas, tanto para os treinos como para os testes, podendo assim verificar se a árvore está ou não *overfitted*.

Algoritmo 2 – Pseudocódigo dos laços de repetição para execução das árvores de decisão

```

1  # Declaracao de todas as opcoes de configuracoes selecionadas
2  test_sizes = [0.1, 0.2, 0.3, 0.4, 0.5]
3  min_samples_splits = [0.1, 0.2, 0.3, 0.4, 0.5]
4  min_samples_leafs = [0.1, 0.2, 0.3, 0.4, 0.5]
5  depths = [2, 3, 5, 8, 13, 21, 34]
6  random_states = [None, 1]
7  criterions = ['gini', 'entropy']
8  splitters = ['random', 'best']
9
10
11 Para cada elemento em test_sizes:
12     Criar variaveis de treino e de teste dividindo os dados da entrada.
13
14     Para cada elemento em min_samples_splits:
15         Para cada elemento em min_samples_leafs:
16             Para cada elemento em depths:
17                 Para cada elemento em random_states:
18                     Para cada elemento em criterions:
19                         Para cada elemento em splitters:
20                             Iniciar o modelo DecisionTreeClassifier com as
                                configuracoes da iteracao
21
22                             Realizar o fit do modelo
23                             Realizar predicoes de treino
24                             Realizar predicoes de teste
25
26                             Criar variaveis com lista de colunas da entrada e
                                metricas do modelo
27
28                             Executar chamada SQL inserindo uma linha na tabela com
                                as informacoes de entrada, configuracao e metricas
                                para o modelo da iteracao

```

Fonte: Elaborado pelo autor

Os conjuntos com diferentes combinações de configurações e dos vários dados disponíveis no *Data Warehouse* tiveram seus registros de execução analisados através da criação de consultas *SQL* onde os resultados eram organizados de forma a ordenar quais conjuntos tiveram melhores métricas. Os resultados com melhores acurácias para cada conjunto de dados foram analisados a fim de encontrar a moda das configurações. Os conjuntos de dados, configurações e resultados obtidos foram abordados e explicados no Capítulo 5.

4.3 REDES NEURAIIS ARTIFICIAIS

Para a implementação das redes neurais artificiais utilizando o *Scikit-learn*, utilizaremos a classe *MLPClassifier* que também possui uma série de parâmetros configuráveis. Dessa forma as redes neurais artificiais foram executadas seguindo a mesma metodologia de laço de repetição especificada para as árvores de decisão, testando a maior quantidade de configurações diferentes possíveis em busca de um melhor resultado. A função *train_test_split* também foi utilizada para dividir e preparar

os dados em suas respectivas variáveis já descritas anteriormente.

Uma vez que a classe *MLPClassifier* não possui um atributo de importância para as classes como a *DecisionTreeClassifier*, utilizamos os mesmos subconjuntos de dados para treinar e testar as redes neurais, possibilitando também a comparação entre os resultados obtidos.

No Quadro 2 é possível observar os valores possíveis para cada parâmetro utilizado na construção deste modelo.

Quadro 2 – Parâmetros escolhidos e seus possíveis valores para Rede Neural

Parâmetro	Valores
test_size	[0.1, 0.2, 0.3, 0.4, 0.5]
random_state	[None, 1]
activation	['identity', 'logistic', 'tanh', 'relu']

Dentre os parâmetros, o *random_state*, quando ativado, determina randomicamente os pesos e iniciações dos viés. O *activation* diz respeito a função de ativação utilizadas nas camadas ocultas da rede neural, utilizando uma função matemática diferente para cada opção. O parâmetro *warm_start* também foi definido como verdadeiro, possibilitando que informações já estabelecidas de execuções anteriores fossem reutilizadas durante o laço de repetição. A tolerância, parâmetro *tolerance*, utilizada para gerar os modelos foi de 0.00001, significando que o modelo só conclui sua execução quando o resultado não tem uma melhora de 0.00001 por pelo menos 10 iterações internas do modelo.

Também pode ser observado no Algoritmo 3 o pseudocódigo utilizado para execução do laço de repetição com todas as configurações escolhidas. Novamente, após criar, treinar e prever os resultados, cada iteração gerou uma rede neural com configuração diferente e teve seus parâmetros e métricas armazenadas em uma tabela do banco de dados.

Para as redes neurais artificiais foi construído uma quantidade menor de modelos uma vez que eles requerem uma maior quantidade de recursos computacionais e levam consideravelmente mais tempo para concluir quando comparadas às árvores de decisão. Sendo assim, os modelos passaram pelo laço de repetição contendo todas as combinações possíveis de parâmetros de acordo com as opções definidas anteriormente.

Após isso os três modelos com melhores performances foram executados

repetidas vezes para calcular a média das métricas, uma vez que a cada execução os resultados são levemente diferentes. Os resultados obtidos foram abordados e detalhados no Capítulo 5.

Algoritmo 3 – Pseudocódigo do laço de repetição para execução das redes neurais artificiais

```

1  # Declaracao de todas as opcoes de configuracoes selecionadas
2  test_sizes = [0.1, 0.2, 0.3, 0.4, 0.5]
3  random_states = [None, 1]
4  activations = ['identity', 'logistic', 'tanh', 'relu']
5  warm_start = True
6  tolerance = 0.00001
7
8
9  Para cada elemento em test_sizes:
10     Criar variaveis de treino e de teste dividindo os dados da entrada.
11
12     Para cada elemento em random_states:
13         Para cada elemento em activations:
14             Iniciar o modelo MLPClassifier com as configuracoes da iteracao
15
16             Realizar o fit do modelo
17             Realizar predicoes de treino
18             Realizar predicoes de teste
19
20
21     Criar variaveis com lista de colunas da entrada e metricas do modelo
22
23     Executar chamada SQL inserindo uma linha na tabela com as informacoes de
        entrada, configuracao e metricas para o modelo da iteracao

```

Fonte: Elaborado pelo autor

5 RESULTADOS E DISCUSSÕES

Nesse capítulo são apresentados os resultados obtidos através da execução dos modelos abordados nos capítulos anteriores. São especificados quais conjuntos de dados e configurações tiveram melhores métricas, assim como quais dados se destacaram e tiveram mais relevância dentro dos modelos.

As métricas utilizadas para avaliar os resultados são a acurácia que representa a fração de predições corretas no geral, a precisão que representa a fração de casos corretamente previstos pelo número total de previsões e o *recall* que representa a fração de eventos positivos que foram previstos corretamente.

5.1 ÁRVORES DE DECISÃO

Para as árvores de decisão, os três conjuntos com melhores métricas foram formados pelas combinações dos dados de carga horária mínima, data de início e data de fim prevista do curso. Estes obtiveram acurácias de aproximadamente 80% e precisão entre 64% e 75%.

As configurações e métricas obtidas são apresentadas no Quadro 3 e Quadro 4 respectivamente.

Quadro 3 – Parâmetros com melhores performances para as árvores de decisão

Conjunto de dados	test_size	criterion	splitter	max_depth	random_state	min_samples_split	min_samples_leaf
1	0.1	entropy	best	13	1	0.1	0.1
2	0.2	entropy	best	13	1	0.1	0.1
3	0.3	entropy	best	13	1	0.1	0.1

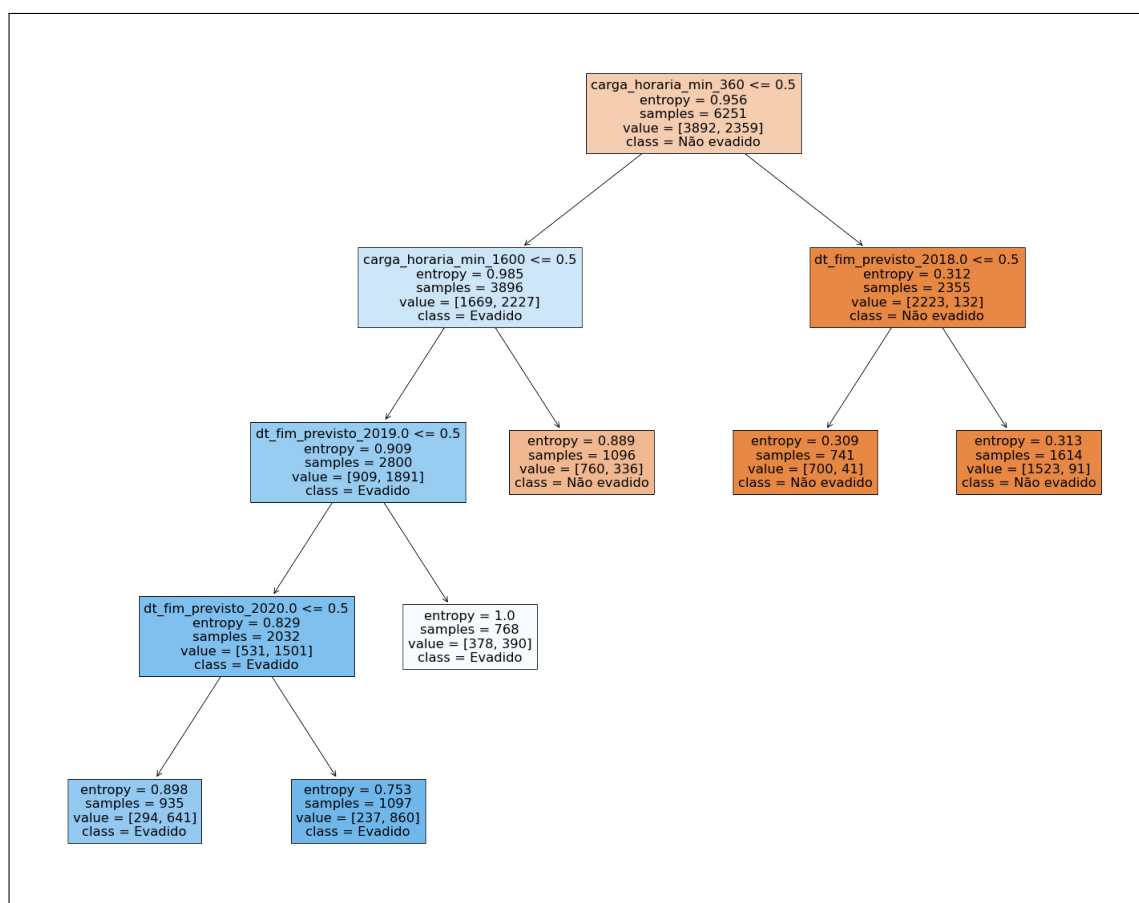
No Quadro 3 o conjunto de dados número 1 é composto apenas pelos dados de carga horária mínima, o conjunto número 2 é formado pelos dados de carga horária mínima e data de fim prevista, já o conjunto de número 3 é composto pelos dados de carga horária mínima, data de início do curso e data de fim prevista. Ao observar o quadro é possível perceber que apenas a configuração *test_size* é diferente para cada conjunto, mostrando que os três tiveram suas melhores acurácias utilizando as mesmas configurações.

Quadro 4 – Média dos resultados para as árvores de decisão

TOP	Acurácia	Precisão	Recall	Número de folhas
1	0.80	0.75	0.826	5
2	0.80	0.643	0.827	6
3	0.798	0.655	0.586	6

No Quadro 4 é possível observar as métricas detalhadas para os três grupos com melhores métricas, todos tiveram uma acurácia próxima e uma precisão com uma variação de aproximadamente 10%. Apenas o *recall* do conjunto de dados número 3 que teve um valor menor que os apresentados nos outros dois conjuntos. O número de folhas dos últimos nós também foi bem próximo para os três casos.

Figura 9 – Árvore de decisão com os dados de carga horária mínima e data de fim previsto



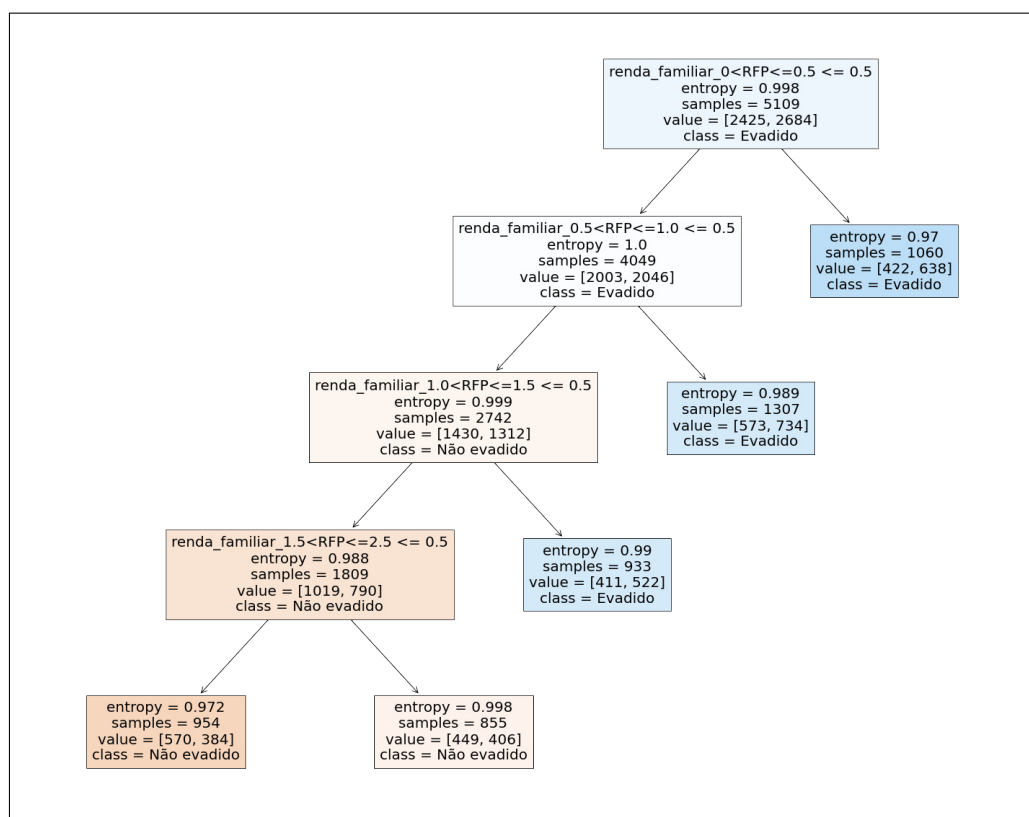
Fonte: Elaborado pelo autor

Foram geradas algumas representações gráficas das árvores de decisão para melhorar a visualização de quais dados específicos tiveram maior importância e influência na classificação de um registro como evadido ou não. Essas representações foram geradas tanto para conjuntos de dados com melhores métricas quanto para outros dois conjuntos de dados que, apesar de não possuírem as maiores métricas, destacam características relevantes na realidade dos alunos.

Através da árvore presente na Figura 9 é possível visualizar que cursos com menor carga horária mínima, entre 360 e 1600 horas, possuem um menor índice de evasão. Enquanto os cursos com carga horária mínima maior, entre 2000 e 2400 horas, indicam uma maior taxa de evasão dos alunos. Também é possível afirmar que alunos que possuem pelo menos 2 anos de curso restante são mais propensos à evasão.

Na Figura 10 é exibida a árvore de decisão gerada com os dados de renda familiar dos alunos. Alunos com renda familiar per capita entre 0 e 1.5 salários mínimos possuem uma maior tendência a evasão, enquanto alunos com rendas maiores que 1.5 salários mínimos geralmente não evadem.

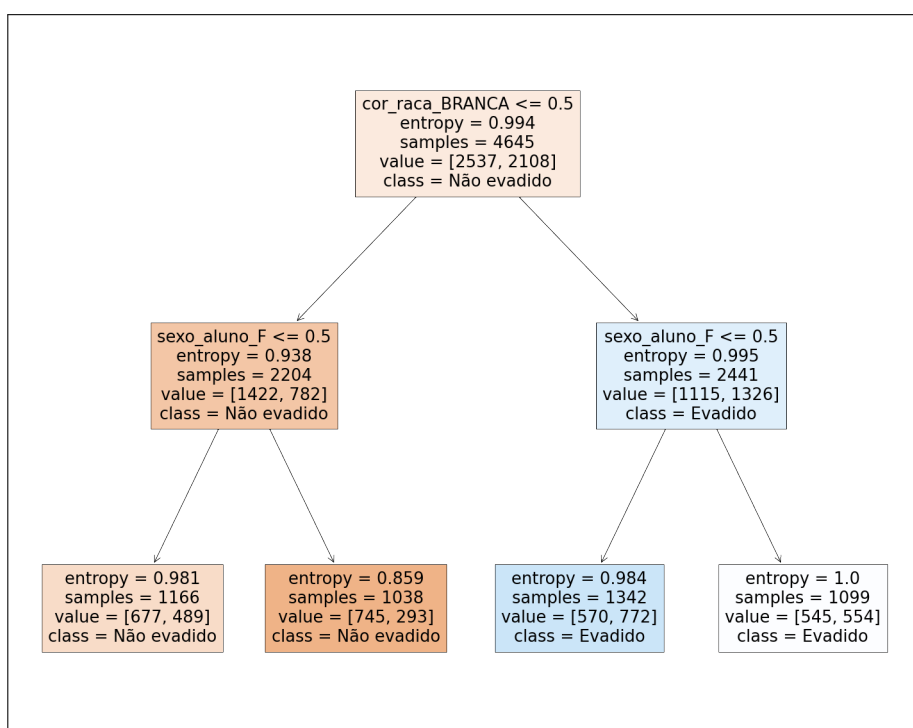
Figura 10 – Árvore de decisão com dados de renda familiar



Fonte: Elaborado pelo autor

Já na Figura 11 é mostrado que alunos autodeclarados como brancos possuem tendência à evasão bem menor do que alunos de outras raças. Alunas do sexo feminino de raças não branca também mostraram grande tendência à evasão.

Figura 11 – Árvore de decisão com dados de sexo e raça



Fonte: Elaborado pelo autor

5.2 REDES NEURAIAS ARTIFICIAIS

Usando os mesmos conjuntos de dados utilizados para as árvores de decisão, as redes neurais artificiais foram construídas e testadas repetidamente até que as configurações e conjuntos com melhores métricas fossem identificados. As configurações escolhidas para os três casos com melhores performances estão descritas no Quadro 5.

Quadro 5 – Parâmetros com melhores performances para as redes neurais artificiais

TOP	test_size	activation	random_state
1	0.4	relu	None
2	0.1	tanh	1
3	0.2	logistic	1

Entre os conjuntos de dados, as informações que mais se destacaram foram as de carga horária mínima, data de início do curso, data de fim prevista e fator de esforço do curso. No Quadro 5, o conjunto de dados número 1 é composto pela combinação da carga horária mínima, fator de esforço e data de início do curso. O conjunto número 2 é formado pela carga horária mínima e data de fim prevista. Já o conjunto número 3 é formado pela carga horária mínima, data de fim prevista e data de início do curso.

Quadro 6 – Média dos resultados para as redes neurais artificiais

Conjunto	Acurácia	Precisão	<i>Recall</i>	Iterações
1	0.829	0.683	0.791	99
2	0.825	0.688	0.72	61
3	0.812	0.675	0.70	171

No Quadro 6 é possível observar as médias das métricas obtidas após a execução repetida das configurações escolhidas como melhores. É possível notar que os três conjuntos tiveram métricas bem parecidas apesar de utilizarem parâmetros de configuração diferentes. Os índices de acurácias e precisão elevados confirmam a relevância dos dados presentes nesses conjuntos.

O conjunto número 1 e 2 obtiveram acurácias e precisões muito próximas de 82% e 68% respectivamente. Os três conjuntos possuem a informação de carga horária mínima e esta se mostrou muito influente ao determinar se um registro seria considerado evadido ou não.

Outros conjuntos de dados que não ficaram entre os três melhores também registraram métricas elevadas, conjuntos formados pelo fator esforço e tipo de curso ou data de fim previsto e tipo de curso obtiveram acurácias próximas de 80%. Os modelos construídos com os dados dos perfis de aluno (raça, sexo e renda familiar) obtiveram acurácias entre 56% e 60%.

6 CONCLUSÃO

Motivado pelo estudo da evasão na rede federal brasileira, a reorganização das informações da Plataforma Nilo Peçanha em um *Data Warehouse* possibilitou a execução de consultas que envolviam muitos dados com facilidade e agilidade. A nova modelagem dimensional proporcionou uma facilidade de leitura para o usuário final assim como o uso de poucas junções entre tabelas para conseguir as informações desejadas. Consultas com centenas de milhares de registros de resposta foram realizadas em poucos segundos, diferentes perspectivas e agrupamentos dos dados foram rapidamente acessados com simples consultas SQL.

Os resultados alcançados com a execução dos algoritmos de mineração obtiveram ótimas métricas e confirmaram perspectivas já divulgadas em outros estudos. Alunos com baixa renda familiar e que não sejam brancos possuem uma tendência mais alta a evadir das instituições de ensino, assim como é afirmado em (UNICEF, 2021). A existência cada vez mais escassa de auxílios financeiros para alunos de baixa renda nas instituições de ensino unida ao racismo estrutural impactam diretamente a permanência de pessoas que venham de realidades mais difíceis ou façam parte de grupos marginalizados da sociedade.

Tanto para as árvores de decisão quanto para as redes neurais artificiais, as informações de carga horária mínima e data de fim prevista foram as que tiveram mais destaque e influência ao predizer se um registro seria classificado como evadido ou não. Cursos com cargas horárias mínimas mais altas e alunos com o fim do curso previsto para dois ou mais anos implicaram em maiores taxas de evasão, mostrando que durante o começo do curso é mais provável que um aluno com um perfil propenso a evasão de fato evada.

Projetos para identificação e auxílio dos alunos predispostos a evasão são de grande necessidade dentro das instituições de ensino. Auxílios financeiros para alunos de grupos marginalizados e baixa renda, além de programas para integrá-los em atividades acadêmicas desde o começo do curso fazem uma grande diferença na realidade do aluno para que ela permaneça na instituição até a conclusão dos seus estudos.

Mesmo com a utilização de um escopo grande como o de matrículas em cursos superiores do ano de 2017, esse trabalho identificou características importantes da realidade dos alunos que impactaram fortemente na classificação destes como evadidos.

É recomendada a realização de trabalhos futuros com análises de escopos menores com o objetivo de encontrar características ainda mais específicas da realidade de cada instituição, campus ou curso. A criação de projetos de pesquisa dedicados a análise desses dados devem gerar resultados importantes para a criação de atividades mais específicas na tentativa de manter os alunos nas instituições.

REFERÊNCIAS

- ARORA, Y. **Building a PostgreSQL Data Warehouse: A Comprehensive Guide**. 2022. Url<https://hevodata.com/learn/postgresql-data-warehouse/>.
- CARVALHO, A. **Redes Neurais Artificiais**. 2009. Url<https://sites.icmc.usp.br/andre/research/neural/>.
- DEVI, R.; NIRMALA, K. Construction of decision tree: Attribute selection measures. **International Journal of Advancements in Research & Technology**, v. 2, n. 4, p. 343–347, 2013.
- DUTRA, J.; SOUZA, J. P.; FERNANDES, D. Classificação de estudantes com potencial à evasão: aplicando mineração de dados no contexto de cursos técnicos subsequentes do ifpb. **Revista Principia - Divulgação Científica e Tecnológica do IFPB**, v. 0, n. 0, 2021. ISSN 2447-9187. Disponível em: <<https://periodicos.ifpb.edu.br/index.php/principia/article/view/5488>>.
- FILHO, R. L. L. S.; MOTEJUNAS, P. R.; HIPÓLITO, O.; LOBO, M. B. d. C. M. A evasão no ensino superior brasileiro. **Cadernos de pesquisa**, SciELO Brasil, v. 37, p. 641–659, 2007.
- HOKAMA, D. D. B.; CAMARGO, D.; FUJITA, F.; FOGLIENE, J. L. V.; JL, V. A modelagem de dados no ambiente data warehouse. **São Paulo**, v. 32, 2004. Disponível em: <<http://meusite.mackenzie.com.br/rogerio/tgi/2004ModelagemDW.pdf>>.
- INESC. Relatório institucional. 2021. Disponível em: <https://www.inesc.org.br/wp-content/uploads/2022/05/Relatorio-institucional-2021_V2.pdf>.
- JOHANN, C. C. et al. Evasão escolar no instituto federal sul-rio-grandense: um estudo de caso no campus passo fundo. Educação, 2012.
- KIMBALL, R.; CASERTA, J. **The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data**. Wiley, 2004. (ITPro collection). ISBN 9780764567575. Disponível em: <<https://books.google.com.br/books?id=ypKilM4fZGIC>>.
- KIMBALL, R.; ROSS, M. The data warehouse toolkit: The definitive guide to dimensional modeling. In: _____. 3. ed. Indianapolis, IN: Wiley, 2013. p. 8. ISBN 978-1-118-53080-1. Disponível em: <<https://www.safaribooksonline.com/library/view/the-data-warehouse/9781118530801/>>.
- MEC; SETEC; DDR. Guia de referência metodológica – pnp 2020. 2021. Disponível em: <http://dadosabertos.mec.gov.br/images/conteudo/pnp/2020/grm_pnp_2020.pdf>.
- NAPOLEÃO, B. **Árvores Decisórias**. 2018. Url<https://ferramentasdaqualidade.org/arvores-decisorias/>.
- NINGNING, G. Proposing data warehouse and data mining in teaching management research. In: **2010 International Forum on Information Technology and Applications**. [S.l.: s.n.], 2010. v. 1, p. 436–439.

SESU/MEC ANDIFES, A. Diplomacao, retencao e evasao nos cursos de graduacao em instituicoes de ensino superior publicas. relatorio da comissao especial de estudos sobre evasao nas universidades publicas brasileiras. 1997. Disponível em: <<https://www.lume.ufrgs.br/bitstream/handle/10183/225423/000224712.pdf?sequence=1>>.

SOUZA, J. P.; FERNANDES, D. Y.; DUTRA, J. Predição precoce de problemas de desempenho de estudantes em modalidade de educação online: um estudo de caso no ensino médio integrado. **Revista Principia - Divulgação Científica e Tecnológica do IFPB**, v. 0, n. 0, 2021. ISSN 2447-9187. Disponível em: <<https://periodicos.ifpb.edu.br/index.php/principia/article/view/5182>>.

TEAM, D. S. **A explicação simples do conceito de retropropagação**. 2020. Url<https://datascience.eu/pt/inteligencia-artificial/como-funciona-o-algoritmo-de-retropropagacao/>.

UNICEF. Cenário da exclusão escolar no brasil. 2021. Disponível em: <<https://www.unicef.org/brazil/media/14026/file/cenario-da-exclusao-escolar-no-brasil.pdf>>.

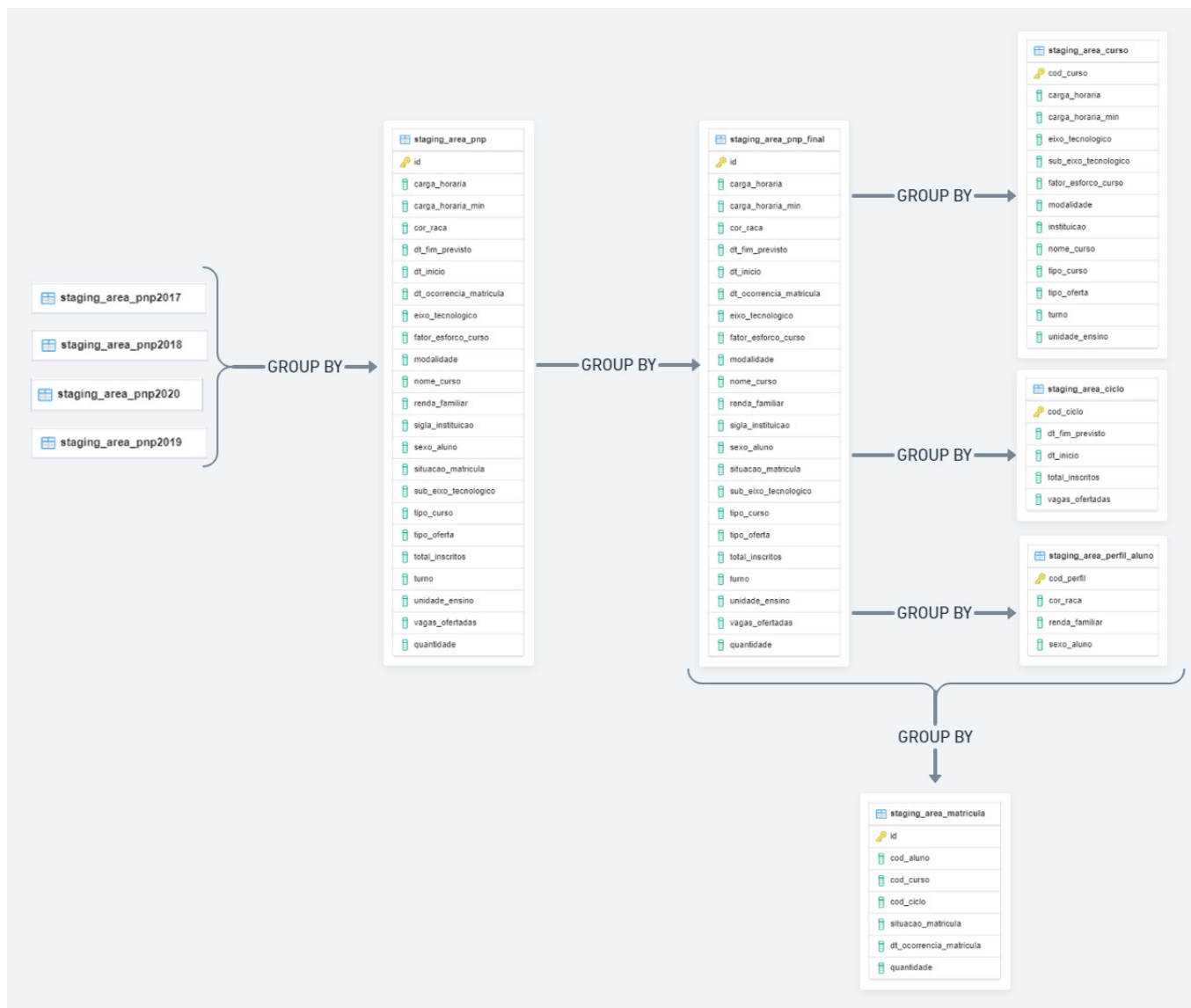
WITTEN, I.; FRANK, E.; HALL, M. **Data Mining: Practical Machine Learning Tools and Techniques**. Elsevier Science, 2011. (The Morgan Kaufmann Series in Data Management Systems). ISBN 9780080890364. Disponível em: <<https://books.google.com.br/books?id=bDtLM8CODsQC>>.

APÊNDICE A – ALGORITMO DE TRANSFORMAÇÃO PARA 2017

```
1 INSERT INTO public.staging_area_pnp(carga_horaria, carga_horaria_min, cor_raca,
  dt_fim_previsto, dt_inicio, dt_ocorrencia_matricula, eixo_tecnologico,
  fator_esforco_curso, modalidade, nome_curso, renda_familiar, sigla_instituicao,
  sexo_aluno, situacao_matricula, sub_eixo_tecnologico, tipo_curso, tipo_oferta,
  total_inscritos, turno, unidade_ensino, vagas_ofertadas, quantidade)
2 (SELECT
3   CARGA_HORARIA,
4   CARGA_HORARIA_MIN,
5   UNACCENT(UPPER(cor_raca)) as COR_RACA,
6   DT_FIM_PREVISTO,
7   DT_INICIO,
8   DT_OCORRENCIA_MATRICULA,
9   UNACCENT(UPPER(eixo_tecnologico)) as EIXO_TECNOLOGICO,
10  REPLACE(fator_esforco_curso, ',', '. ') as FATOR_ESFORCO_CURSO,
11  UNACCENT(UPPER(modalidade)) as MODALIDADE,
12  UNACCENT(UPPER(nome_curso)) as NOME_CURSO,
13  UPPER(REPLACE(renda_familiar, ',', '. ')) as RENDA_FAMILIAR,
14  UNACCENT(UPPER(sigla_instituicao)) as SIGLA_INSTITUICAO,
15  UPPER(LEFT(sexo_aluno,1)) as SEXO_ALUNO,
16  UNACCENT(REPLACE(UPPER(situacao_matricula), '_ ', ' ')) as SITUACAO_MATRICULA,
17  UNACCENT(UPPER(sub_eixo_tecnologico)) as SUB_EIXO_TECNOLOGICO,
18  UNACCENT(UPPER(tipo_curso)) as TIPO_CURSO,
19  UNACCENT(UPPER(tipo_oferta)) as TIPO_OFERTA,
20  SUM(total_inscritos::INTEGER) as TOTAL_INSCRITOS,
21  UNACCENT(UPPER(turno)) as TURNO,
22  UNACCENT(UPPER(unidade_ensino)) as UNIDADE_ENSINO,
23  SUM(split_part(vagas_ofertadas, ',', 1)::INTEGER) as VAGAS_OFERTADAS,
24  COUNT(*) QUANTIDADE
25 FROM
26   STAGING_AREA_PNP2017
27 GROUP BY
28   CARGA_HORARIA, CARGA_HORARIA_MIN,
29   UNACCENT(UPPER(cor_raca)), DT_FIM_PREVISTO,
30   DT_INICIO,
31   DT_OCORRENCIA_MATRICULA,
32   UNACCENT(UPPER(eixo_tecnologico)),
33   REPLACE(fator_esforco_curso, ',', '. '),
34   UNACCENT(UPPER(modalidade)),
35   UNACCENT(UPPER(nome_curso)),
36   UPPER(REPLACE(renda_familiar, ',', '. ')),
37   UNACCENT(UPPER(sigla_instituicao)),
38   UPPER(LEFT(sexo_aluno,1)),
39   UNACCENT(REPLACE(UPPER(situacao_matricula), '_ ', ' ')),
40   UNACCENT(UPPER(sub_eixo_tecnologico)),
41   UNACCENT(UPPER(tipo_curso)),
42   UNACCENT(UPPER(tipo_oferta)),
43   UNACCENT(UPPER(turno)),
44   UNACCENT(UPPER(unidade_ensino));
```

Fonte: Elaborado pelo autor

APÊNDICE B – FLUXO DE TRANSFORMAÇÃO DOS DADOS DA PNP



ANEXO A – DICIONÁRIO DE DADOS PNP 2017

Microdados

Os microdados utilizados na Plataforma Nilo Peçanha – PNP estão disponíveis para consulta e download. Os arquivos estão estruturados e organizados no formato CSV e são compostos pelo menor nível de fragmentação de dados recolhidos pela Plataforma.

Os microdados da PNP estão dispostos em quatro arquivos:

- Microdados Matrículas;
- Microdados Eficiência Acadêmica;
- Microdados Servidores;
- Microdados Financeiros.

Microdados Matrículas

Representa o conjunto de dados relacionados às matrículas (matrícula, aluno, ciclo de matrícula, curso, unidade de ensino, instituição). Possui uma linha para cada matrícula que esteve ativa em pelo menos um dia no ano base 2017.

A tabela é composta pelas seguintes informações:

Coluna	Descrição
Carga Horaria	Carga horária do ciclo de matrícula declarada no SISTEC e validada na PNP pela instituição.
Carga Horaria Minima	Carga horária mínima do curso de acordo com o art. 5º, §2º, da Portaria SETEC nº 25, de 13/08/2015.
Co Matricula	Código da matrícula no SISTEC.
Cor Raca	Cor/Raça do aluno declarado na PNP pela instituição. Opções: Amarela, Branca, Indígena, Parda, Preta e Não Declarada.
Dt Data Fim Previsto	Data prevista para o final do ciclo de matrícula validado/atualizado pela instituição na PNP.
Dt Data Inicio	Data de início do ciclo de matrícula validado/atualizado pela instituição na PNP.
Dt Ocorrencia Matricula	Mês/Ano em que a matrícula foi efetivada na instituição. Não se refere à data em que a matrícula foi cadastrada no sistema. O campo foi validado/atualizado pela instituição na PNP. A referência para data é sempre o dia 1º.
Eixo Tecnologico	Eixo tecnológico do curso associado pela instituição na PNP.
Fator Esforco Curso	Índice de ajuste da carga horária do curso em função da quantidade de aulas práticas com redução do número de alunos em decorrência da subdivisão da turma, conforme valores relacionados no Anexo I da Portaria SETEC nº 25, de 13/08/2015 e o nome de curso associado na PNP pela instituição.
Mes De Ocorrencia	Mês/Ano em que a situação da matrícula efetivamente mudou. Não se refere à data em que a situação da matrícula foi atualizada no sistema. Para as matrículas que não tiveram situação alterada,

Coluna	Descrição
	corresponde ao mês de ocorrência da matrícula. A referência para data é sempre o dia 1º. Campo validado/atualizado pela instituição na PNP.
Modalidade Ensino	Classificação para identificar ensino presencial ou ensino a distância.
Nome Curso	Nomenclatura adotada para padronização dos nomes de cursos de acordo com: Resoluções CNE, CNCT, CNCST, Guia FIC, dentre outros. Os nomes dos cursos cadastrados no SISTEC foram associados às opções disponíveis na PNP. Aqueles que não possuíam correspondência foram associados à nomenclatura composta pelo “Tipo de Curso – Eixo Tecnológico”.
Fonte de financiamento	Informa se a matrícula é financiada por recursos próprios (Sem Programa Associado) ou pelos Programas UAB ou PRONATEC - Rede e-Tec Brasil.
Renda Familiar	Faixa de renda per capita familiar do aluno, declara na PNP pela instituição (Opções: $0 < RFP \leq 0,5$; $0,5 < RFP \leq 1$; $1 < RFP \leq 1,5$; $1,5 < RFP \leq 2,5$; $2,5 < RFP \leq 3,5$; $RFP > 3,5$; Não declarada).
Sg Inst	Sigla da Instituição na PNP. Cada escola técnica vinculada às Universidades Federais foi compreendida como uma Instituição.
Sg Sexo	Informa o sexo do estudante constante no SISTEC.
Situa__O_Matricula	Atributo que apresenta a situação da matrícula do aluno conforme declarado/validado na PNP. A situação inicial da matrícula é “Em Curso”. As situações possíveis são: . Em Curso: matrícula ativa. . Cancelada: aplica-se ao aluno que solicita formalmente o cancelamento da sua matrícula antes de iniciar as atividades pedagógicas do curso. . Abandono: aplica-se ao aluno que possui mais de 25% de faltas não justificadas. Recomenda-se se modificar o status para “Abandono” somente quando não houver mais possibilidade de o aluno voltar a frequentar as aulas. . Concluída: aplica-se ao aluno que concluiu todo o curso com êxito. . Desligada: aplica-se ao aluno que solicita formalmente o cancelamento da sua matrícula após iniciar as atividades pedagógicas do curso. . Integralizada: aplica-se ao aluno que concluiu a parte teórica do curso, mas está devendo o estágio obrigatório, TCC, monografia, dissertação ou tese. . Reprovada: O aluno finalizou o curso porém não logrou êxito nas avaliações. Aplica-se nos casos de impossibilidade de continuação do curso. . Transf_ext: aplica-se ao aluno que será transferido para outra Unidade de Ensino. . Transf_int: aplica-se ao aluno que muda de curso, dentro da mesma Unidade de Ensino. O campo foi validado/atualizado pela instituição na PNP.
Sub Eixo Tecnológico	Categorização própria da PNP para distinguir cursos de um mesmo Eixo Tecnológico em suas diferentes áreas de concentração.

Coluna	Descrição
Tipo Curso	Categorização transversal utilizada para diferenciar os cursos da EPCT em seus diferentes níveis e graus. Opções: Educação Infantil, Ensino Fundamental I, Ensino Fundamental II, Ensino Médio, Qualificação Profissional (FIC), Técnico, Tecnologia, Licenciatura, Bacharelado, Especialização (Lato Sensu), Mestrado Profissional, Mestrado, Doutorado Profissional e Doutorado.
Tipo Oferta	Categorização transversal utilizada para diferenciar as formas de oferta dos Cursos Técnicos e de Qualificação Profissional (FIC). Opções: Integrado, Subsequente, Concomitante, PROEJA – Concomitante e PROEJA – Integrado.
Total Inscritos	Corresponde aos candidatos participantes dos processos seletivos que concorreram às vagas disponibilizadas para a fase inicial de um ciclo de matrícula em determinado curso, em suas diversas formas de seleção.
Turno	período de tempo determinado em que o aluno cursa a maior parte das aulas. Opções: matutino, vespertino, noturno ou integral. Não se aplica aos cursos com Modalidade de Ensino a distância.
Unidade Ensino	Nome da unidade de ensino a qual a matrícula está vinculada.
Vagas Ofertadas	Corresponde ao número de vagas disponibilizadas para novas matrículas no início do ciclo de um curso, considerando todas as formas de ingresso disponibilizados (vestibular, sorteio, SISU ou outras formas de ingresso) no ano de referência.

Microdados Eficiência Acadêmica

Representa o conjunto de dados necessários para o cálculo dos Indicadores disponíveis no painel “Eficiência Acadêmica”. Assim como os microdados citados acima, este arquivo também está relacionado às matrículas (matrícula, aluno, ciclo de matrícula, curso, unidade de ensino, instituição), contudo possui uma linha para cada matrícula de ciclos que estavam previstos para terminar no ano anterior ao ano de referência.

A tabela é composta pelas seguintes informações:

Coluna	Descrição
Carga Horaria	Carga horária do ciclo de matrícula declarada no SISTEC e validada na PNP pela instituição.
Carga Horaria Mínima	Carga horária mínima do curso de acordo com o art. 5º, §2º, da Portaria SETEC nº 25, de 13/08/2015.
Co Matrícula	Código da matrícula no SISTEC.
Dt Data Fim Previsto	Data prevista para o final do ciclo de matrícula validado/atualizado pela instituição na PNP.
Dt Data Inicio	Data de início do ciclo de matrícula validado/atualizado pela instituição na PNP.
Dt Ocorrência Matrícula	Mês/Ano em que a matrícula foi efetivada na instituição. Não se refere à data em que a matrícula foi cadastrada no sistema. O campo foi validado/atualizado pela instituição na PNP. A referência para data é sempre o dia 1º.

Coluna	Descrição
Eixo Tecnológico	Eixo tecnológico do curso associado pela instituição na PNP.
Fator Esforço Curso	Índice de ajuste da carga horária do curso em função da quantidade de aulas práticas com redução do número de alunos em decorrência da subdivisão da turma, conforme valores relacionados no Anexo I da Portaria SETEC nº 25, de 13/08/2015 e o nome de curso associado na PNP pela instituição.
Mês De Ocorrência	Mês/Ano em que a situação da matrícula efetivamente mudou. Não se refere à data em que a situação da matrícula foi atualizada no sistema. Para as matrículas que não tiveram situação alterada, corresponde ao mês de ocorrência da matrícula. A referência para data é sempre o dia 1º. Campo validado/atualizado pela instituição na PNP.
Modalidade Ensino	Classificação para identificar ensino presencial ou ensino a distância.
Nome Curso	Nomenclatura adotada para padronização dos nomes de cursos de acordo com: Resoluções CNE, CNCT, CNCST, Guia FIC, dentre outros. Os nomes dos cursos cadastrados no SISTEC foram associados às opções disponíveis na PNP. Aqueles que não possuíam correspondência foram associados à nomenclatura composta pelo “Tipo de Curso – Eixo Tecnológico”.
Fonte de financiamento	Informa se a matrícula é financiada por recursos próprios (Sem Programa Associado) ou pelos Programas UAB ou PRONATEC - Rede e-Tec Brasil.
Sg Inst	Sigla da Instituição na PNP. Cada escola técnica vinculada às Universidades Federais foi compreendida como uma Instituição.
Sg Sexo	Informa o sexo do estudante constante no SISTEC.
Situa__O_Matricula	Atributo que apresenta a situação da matrícula do aluno conforme declarado/validado na PNP. A situação inicial da matrícula é “Em Curso”. As situações possíveis são: . Em Curso: matrícula ativa. . Cancelada: aplica-se ao aluno que solicita formalmente o cancelamento da sua matrícula antes de iniciar as atividades pedagógicas do curso. . Abandono: aplica-se ao aluno que possui mais de 25% de faltas não justificadas. Recomenda-se se modificar o status para “Abandono” somente quando não houver mais possibilidade de o aluno voltar a frequentar as aulas. . Concluída: aplica-se ao aluno que concluiu todo o curso com êxito. . Desligada: aplica-se ao aluno que solicita formalmente o cancelamento da sua matrícula após iniciar as atividades pedagógicas do curso. . Integralizada: aplica-se ao aluno que concluiu a parte teórica do curso, mas está devendo o estágio obrigatório, TCC, monografia, dissertação ou tese. . Reprovada: O aluno finalizou o curso porém não logrou êxito nas avaliações. Aplica-se nos casos de impossibilidade de continuação do curso. . Transf_ext: aplica-se ao aluno que será transferido para outra Unidade de Ensino.

Coluna	Descrição
	. Transf_int: aplica-se ao aluno que muda de curso, dentro da mesma Unidade de Ensino. O campo foi validado/atualizado pela instituição na PNP.
Sub Eixo Tecnológico	Categorização própria da PNP para distinguir cursos de um mesmo Eixo Tecnológico em suas diferentes áreas de concentração.
Tipo Curso	Categorização transversal utilizada para diferenciar os cursos da EPCT em seus diferentes níveis e graus. Opções: Educação Infantil, Ensino Fundamental I, Ensino Fundamental II, Ensino Médio, Qualificação Profissional (FIC), Técnico, Tecnologia, Licenciatura, Bacharelado, Especialização (Lato Sensu), Mestrado Profissional, Mestrado, Doutorado Profissional e Doutorado.
Tipo Oferta	Categorização transversal utilizada para diferenciar as formas de oferta dos Cursos Técnicos e de Qualificação Profissional (FIC). Opções: Integrado, Subsequente, Concomitante, PROEJA – Concomitante e PROEJA – Integrado.
Total Inscritos	Corresponde aos candidatos participantes dos processos seletivos que concorreram às vagas disponibilizadas para a fase inicial de um ciclo de matrícula em determinado curso, em suas diversas formas de seleção.
Unidade Ensino	Nome da unidade de ensino a qual a matrícula está vinculada.
Vagas Ofertadas	Corresponde ao número de vagas disponibilizadas para novas matrículas no início do ciclo de um curso, considerando todas as formas de ingresso disponibilizados (vestibular, sorteio, SISU ou outras formas de ingresso) no ano de referência.

Microdados Servidores

Representa o conjunto de dados relacionados aos servidores docentes e técnicos-administrativos.

A tabela é composta pelas seguintes informações:

Coluna	Descrição
Número do registro	Numeração sequencial neste arquivo.
Sg Instituto	Sigla da Instituição na PNP. Cada escola técnica vinculada às Universidades Federais foi compreendida como uma Instituição.
Nome Unidade	Nome da unidade de ensino a qual o servidor está vinculado.
Cod Unidade	Código da Unidade Gestora (UG) no SIAPE. A unidade de ensino pode possuir mais de uma UG. O vínculo entre a UG e a Unidade de Ensino foi realizado na Plataforma pela Revalide.
Vínculo Contrato	Refere-se a forma de contrato do servidor cadastrado no SIAPE: Efetivo ou Substituto/Temporário
Vínculo Carreira	Aponta a qual carreira o servidor pertence: EBTT, PCCTAE ou PROF. OUTROS.
Vínculo Classe	Nível de classificação da carreira.

Coluna	Descrição
Vinculo Jornada	Indica a jornada de trabalho do servidor.
Formacao Servidor	Aponta a titulação do servidor.
Rsc Servidor	Indica se o servidor possui RSC (Reconhecimentos do Saberes e Competências)
Vinculo Professor	Indica se o servidor é ou não docente.

Microdados Financeiros

Representa o conjunto de dados para as variáveis relacionadas ao financeiro.

A tabela é composta pelas seguintes informações:

Coluna	Descrição
UO	Entidade da administração direta, inclusive fundo ou órgão autônomo, da administração indireta (autarquia, fundação ou empresa estatal) em cujo nome a lei orçamentária ou crédito adicional consigna, expressamente, dotações com vistas à sua manutenção e à realização de um determinado programa de trabalho. Constituem desdobramentos dos órgãos orçamentários.
Ação	Projeto, atividade ou operação especial em que um programa está detalhado. A ação é definida por descrição e código de quatro dígitos.
GND	Classificação da despesa agregando elementos de despesa com as mesmas características quanto ao objeto de gasto, conforme discriminado a seguir: 1 - Pessoal e Encargos Sociais; 3 - Outras Despesas Correntes; 4 - Investimentos; 5 - Inversões Financeiras. Foi incluído também nesta coluna o termo "33914712 CONTRIBUICAO P/ O PIS/PASEP".
Despesas Liquidadas	É o valor que foi disponibilizado para pagamento efetivo após a verificação do cumprimento contratual.
Restos a Pagar Não-Processados Pagos	É o valor que foi empenhado em anos anteriores, contudo liquidado e pago no ano de referência.