



MATEUS DO NASCIMENTO MAGALHÃES DA SILVA
FABRICIO ALMEIDA DA SILVA NUNES

**BIG DATA ANALYTICS E DATA-DRIVEN-DESIGN NA EDUCAÇÃO: ANÁLISE DO
DESEMPENHO E DOS DADOS SOCIOECONÔMICOS DOS PARTICIPANTES DO
ENEM 2019 ATRAVÉS DA MINERAÇÃO DE DADOS E BUSINESS INTELLIGENCE**

RIO DE JANEIRO

2021

MATEUS DO NASCIMENTO MAGALHÃES DA SILVA

FABRICIO ALMEIDA DA SILVA NUNES

BIG DATA ANALYTICS E DATA-DRIVEN-DESIGN NA EDUCAÇÃO: ANÁLISE DO DESEMPENHO E DOS DADOS SOCIOECONÔMICOS DOS PARTICIPANTES DO ENEM 2019 ATRAVÉS DA MINERAÇÃO DE DADOS E BUSINESS INTELLIGENCE

Trabalho de conclusão de curso apresentado ao Centro Universitário Carioca (Unicarioca), como requisito exigido parcial à obtenção do grau de Bacharel em Ciência da Computação.

Orientador(a): Prof.^a Daisy Cristine Albuquerque da Silva, M.Sc.

RIO DE JANEIRO

2021

MATEUS DO NASCIMENTO MAGALHÃES DA SILVA
FABRICIO ALMEIDA DA SILVA NUNES

TEMA: Data Science na Educação: Extração de Conhecimento Sobre Desempenho dos Participantes do Exame Nacional do Ensino Médio a partir de Mineração de Dados

Banca Examinadora

AGRADECIMENTOS

À Deus, pois sem ele nada é possível.

À nossa orientadora, por toda ajuda e paciência durante o desenvolvimento deste trabalho.

À Unicarioca, seu corpo docente e funcionários.

DEDICATÓRIA

RESUMO

O presente trabalho utiliza a arquitetura moderna de business intelligence e engenharia de grandes volumes de dados para realizar análise de dados ad-hoc e técnicas de mineração de dados com foco no levantamento estatístico de dados sociais, econômicos e de desempenho dos participantes do exame nacional do ensino médio (ENEM) no ano de 2019. O fluxo de ETL e a modelagem de um relational data mart permitiram aplicar o algoritmo de descoberta de regras de associação para definir a correlação e o grau probabilístico de causalidade entre variáveis correspondentes a informações pessoais, as provenientes do questionário socioeconômico e as que correspondem ao desempenho dos participantes nas cinco áreas de conhecimento e na redação e suas competências. O conjunto de dados modelado também possibilitou a análise de dados ad-hoc para levantamento estatístico do perfil dos participantes.

ABSTRACT

The current study utilizes the modern business intelligence architecture and engineering of big volumes of data to carry ad-hoc data analysis and data mining techniques focused on the statistical survey of social, economic and performance data of the participants of the National High School Exam (ENEM) in 2019. The ETL flow and the modeling of a relational data mart allowed applying the association rule discovery algorithm to define the correlation and the probabilistic degree of causality between variables corresponding to personal information, those from the socioeconomic questionnaire, and those corresponding to the participants performance in the five knowledge areas and in the essay and its competencies. The modeled dataset also enabled ad-hoc data analysis to statistically survey the participant's profile.

LISTA DE ILUSTRAÇÕES

Figura 1 - Gráfico de linha contendo análise de tempo de interesse sobre fogão, ar-condicionado, aquecedor e ferro de passar.....	25
Figura 2 - Gráfico de linha contendo análise de tempo de interesse sobre geladeira, freezer, máquina de lavar louça, secadora de roupas e forno de micro-ondas.	26
Figura 3 - Gráfico de linha contendo análise de tempo de interesse sobre o aspirado de pó e sobre a lavadora de roupas.....	26
Figura 4 - Informações da coluna de status de redação.	28
Figura 5 - Gastos da assinatura gratuita para usuário comum.....	29
Figura 6 - Gastos do usuário 1 da assinatura para estudante.....	29
Figura 7 - Gastos do usuário 2 na assinatura para estudante.....	30
Figura 8 - Gasto dos recursos do Data Lake do usuário 2.....	32
Figura 9 - Gasto dos recursos do SQL do Azure do usuário 1.....	34
Figura 10 - Gastos do Analysis Services, Recurso B1.....	35
Figura 11 - Gastos dos recursos do Data Factory da assinatura de estudante do usuário 1.....	36
Figura 12 - Gastos por recursos.....	36
Figura 13 - Ciência de dados no contexto dos diversos processos relacionados a dados na organização.....	50
Figura 14 - A pirâmide de conhecimento.....	52
Figura 15 - Diagrama Entidade Relacionamento de sistema de imobiliária.....	66
Figura 16 - Esquema Estrela (Star Schema).....	67
Figura 17 - Esquema flocos de neve(Snowflake Schema).....	68
Figura 18 - Diretório de arquivos do Data Lake da Azure.....	89
Figura 19 - Configuração dos pipelines para o Azure SQL DB.....	90
Figura 20 - Congiguração do Data Factory na Azue.....	91
Figura 21 - Integração do MSSQL da Azure.....	91
Figura 22 - Relacionamento entre as tabelasdo banco de dados.....	92
Figura 23 - Conexão do Data Lake com Servidor SQL.....	92
Figura 24 - Relacionamento entra as tabelas do banco de dados transformado.....	93
Figura 25 - Tabela de quantidade de inscritos segmentados pelo item cor e raça.....	94
Figura 26 - Configuração de dados de uma coluna.....	95
Figura 27 - Tabela utilizando filtragem de dados.....	95
Figura 28 - Filtragem de dados de uma coluna no Power BI.....	96

Figura 29 - Tabela de média das notas da redação por faixa de idade	96
Figura 30 - Configuração do Power BI para a definição da operação de média	97
Figura 31 - Tabela de porcentagem das faixas das notas da prova objetiva aplicada a gênero feminino.....	97
Figura 32 - Gráfico e colunas que mostra a comparação de totais de inscritos para cada região do Brasil.....	98
Figura 33 - Mapa de árvore exibindo uma comparação entre a quantidade de inscritos de distintos sexos e suas respectivas faixas notas	99
Figura 34 - Gráfico de cascata que exibi as distribuições da proporção das quantidades de celulares que há na residência dos inscritos	99
Figura 35 - Visualização de funil das informações das faixas de idade com a nota média da redação.....	100
Figura 36 -Gráfico de pizza exibindo a média da nota da redação segmentado pelo estado civil dos participantes do exame.....	101
Figura 37 -Gráfico de rosca exibindo a média da nota da redação segmentado pelo estado civil dos participantes do exame.....	101
Figura 38 - Tabela comparativa sobre DVD e TV por assinatura junto da média da nota da redação.....	102
Figura 39 - média de notas da redação distribuídas por cidades do Brasil.....	102
Figura 40 - Visualização de cartão, destacando o número de inscritos no exame	103
Figura 41 - Segmentação dos gráficos que contêm as informações de nota da redação	104
Figura 42 - Segmentação com a opção trocada e a interação dos visuais	104
Figura 43 - Gráfico de árvore hierárquica com os dados dos participantes solteiros, brancos, brasileiros com faixa de nota entre 400 a 600 pontos na prova objetiva, mensurando pela quantidade de inscrito por atributos.....	105
Figura 44 - Pré-processamento – Tratamento de redundâncias - Estado Civil	109
Figura 45- Pré-processamento – Tratamento de redundâncias - Cor ou raça.....	109
Figura 46 - Pré-processamento – Tratamento de redundâncias - Questionários Socioeconômicos.....	110
Figura 47 - Transformação - Cálculo de média das notas	111
Figura 48 - Transformação - Tratamento de nulos – Idade - 2.....	111
Figura 49 - base de dados transformada - Classificação de discretos – Idade - 1	112
Figura 50 - base de dados transformada - Classificação de discretos – Idade - 2	112

Figura 51 - leitura de transações e remoção de cabeçalhos do arquivo contendo questionários de situação familiar.....	113
Figura 52 - base de dados transformada - 2.....	114
Figura 53 - base de dados transformada - 1.....	114
Figura 54 - base de dados transformada - 2.....	115
Figura 55 - Documentos CSVs de consultas para os casos de uso.....	116
Figura 56 - Leitura e geração de regras dos casos de uso de questionários socioeconômicos	117
Figura 57 - Leitura e geração de regras dos casos de uso de informações pessoais, da prova, do ensino médio e de localidade.....	117
Figura 58 - Alteração da codificação das tabelas	121
Figura 59 - Transformação das colunas sobre o a nacionalidade dos participantes.	122
Figura 60 - Adicionando uma coluna condicional.....	122
Figura 61 - Distribuição por faixa da tabela Info_Pessoal	123
Figura 62 - Criação da coluna condicional para a faixa de idade na tabela Info_Pessoal	123
Figura 63 - Colunas das tabelas de inclusão.....	124
Figura 64 - Adicionando coluna condicional a partir das colunas sobre a utilização de recursos	125
Figura 65 - quantidade de inscritos por faixa da média nota da prova objetiva	126
Figura 66 - Gráfico de colunas onde exibe a quantidade de inscritos por regiões do Brasil, situando as suas respectivas residências	126
Figura 67 - Distribuição da renda familiar pela quantidade de inscritos.....	127
Figura 68 - Gráfico de barras que mostra a comparação dos inscritos com a situação do ensino médio	128
Figura 69 - Gráfico de rosca exibindo a diferença de quantidade de inscritos que solicitaram atendimento específico	128
Figura 70 - Gráfico de árvore hierárquica exibindo o desenvolvimento de um participante através de seus atributos	129
Figura 71 - Gráfico de pareto, para comparar as regiões e as notas dos participantes	130
Figura 72 - Caso de uso - Informações pessoais	131
Figura 73 - Caso de uso - Informações pessoais - 2	131
Figura 74 - Caso de uso - Situação familiar - 1	131
Figura 75 - Caso de uso - Situação familiar - 2	132

Figura 76 - Caso de uso - Situação familiar - 3	132
Figura 77 - Caso de uso - Situação doméstica	132
Figura 78 - Caso de uso - Multimídia e telecomunicação	133
Figura 79 - Caso de uso - Eletrodomésticos	133
Figura 80 - Caso de uso - Situação do ensino médio.....	133
Figura 81 - Caso de uso - Situação do ensino médio - 2	134
Figura 82 - Caso de uso - Informações da prova	134
Figura 83 - Caso de uso - Localidade	134
Figura 84 -Gráfico de pareto exibindo uma comparação entre as rendas familiares e as aplicando as médias da nota redação a quantidade de pessoas por residência.	135
Figura 85 - Gráfico de árvore hierárquica que demonstra o caminho percorrido e aplicando a quantidade de participantes que está correlacionado a cada atributo	136
Figura 86 - Gráfico de árvore hierárquica que demostra o caminho percorrido e aplica a média da redação para cada atributo correlacionado.....	136

LISTA DE TABELAS

Tabela 1 - Classificação dos eletrométricos pontuas através do Google Trends	27
Tabela 2 - Pós Pago	30
Tabela 3 - Pacotes de compromisso mensal	31
Tabela 4 - Transição de dados	31
Tabela 5 - Preço Analysis Services	35

LISTA DE ABRIVEATURA E SIGLAS

AI - Aprendizado indutivo

AM - Aprendizado de Máquina

LS – Linked Server

AAS – Azure Analysis Services

CSV - Comma-separated-values

DAU - Daily Active Users

ETL – Extract Transform Load

KDD – Knowledge Discovery in Databases

DIKW– Data Information Knowledge Wisdom

PAAS – Plataform as a Service

IAAS – Infrastructure as a Service

SAAS – Software as a Service

OLTP – On-line Transaction Processing

OLAP – On-line Analytical Processing

ARM – Association Rule Mining

DRA – Descoberta de Regras de Associação

SMS - Short Message Service

MD- Mineração de Dados

ENEM – Exame Nacional do Ensino Médio

INEP - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira

MEC - Ministério da Educação

IA - Inteligência Artificial

CRAN – Comprehensive R Archive Network

DAX – Data Analysis Expressions

SUMÁRIO

1.	INTRODUÇÃO.....	17
1.1.	MOTIVAÇÃO E JUSTIFICATIVA	18
1.1.1.	Objetivo geral	19
1.1.2.	Objetivos específicos	20
1.2	DIFICULDADES E LIMITAÇÕES DO TRABALHO	21
1.1.3.	Ideia inicial	21
1.1.4.	Limitações tecnológicas.....	22
1.1.5.	Limitações da base de dados.....	23
1.1.6.	Custos de infraestrutura	28
1.3	CONTEXTUALIZAÇÃO DO PROBLEMA	36
1.4	STACK DE TECNOLOGIAS E FERRAMENTAS ESCOLHIDAS	38
1.5	ANONIMIZAÇÃO DE DADOS SENSÍVEIS E LGPD	38
2.	BIG DATA, ANÁLISE DE DADOS E O ENEM	40
2.1	INEP.....	40
2.2	EXAME NACIONAL DO ENSINO MÉDIO.....	40
2.2.1	Acessibilidade e Inclusão social	42
2.3	BIG DATA	43
2.4	Engenharia de dados e Big Data	44
2.4.1	Tipos de processamento de dados em pipelines	45
2.5	BUSINESS INTELLIGENCE.....	46
2.5.1	Arquitetura de BI	46
2.5.2	Data-Driven-Design (DDD)	47
2.6	ANÁLISE DE DADOS	48
2.6.1	Tipos de análise de dados em Big Data	48
2.6.2	Análise de dados Ad Hoc	49

2.7	Ciência de dados	50
2.8	MINERAÇÃO DE DADOS	51
2.8.1	A pirâmide do conhecimento para mineração de dados: Hierarquia DIWK 51	
2.8.2	KDD	53
2.8.3	Técnicas de mineração de dados	55
2.9	ETL & ELT	59
2.9.1	ELT	60
2.10	MODELO DE PROCESSAMENTO DE DADOS EM BANCOS DE DADOS 60	
2.10.1	OLTP	61
2.10.2	OLAP	61
2.11	MODELO DE ARMAZENAMENTO DE DADOS MASSIVOS (BIG DATA) 61	
2.11.1	Data Warehouses	62
2.11.2	Data Mart	62
2.11.3	Data lake	62
2.11.4	Modern Datawarehouse(MDW)	63
2.12	MODELAGEM DE DATABASES E DATAWHAREHOUSES	64
2.12.1	Modelo e Diagrama Entidade-Relacionamento (MER / DER)	64
2.12.2	Modelagem de Star Schema e Snowflake	66
2.13	CUBOS OLAP	68
2.13.1	Multidimensional	68
2.13.2	Tabular	69
2.14	CLOUD SERVICES (SERVIÇO EM NUVENS)	69
2.14.1	SaaS	70
2.14.2	IaaS	71
2.14.3	PaaS	71

2.14.4	DaaS.....	71
2.14.5	CaaS	71
2.14.6	EaaS	72
2.14.7	Serviços escolhidos.....	72
2.15	FORMATOS DE DOCUMENTOS DE TEXTO.....	73
2.16	VISUALIZADORES DE DADOS	74
2.16.1	Visualizador escolhido: Power BI	76
2.16.2	Power Query/M language (Marcação).....	77
2.17	LINGUAGEM PARA ANÁLISE DE DADOS	77
2.17.1	DAX.....	77
2.17.2	R language and environment	78
2.17.3	SQL.....	79
2.18	FERRAMENTAS DE DESENVOLVIMENTO.....	80
2.18.1	Storage Explorer	81
2.18.2	IDEs e DBMS	81
2.18.3	DBMS(SGBD).....	83
3.	PROJETO DE BI MODERNO COM MINERAÇÃO E ANÁLISE DE DADOS PARA ESTUDO DO ENEM 2019	86
3.1	DESCRIÇÃO DO PROCESSO OPERACIONAL.....	86
3.1.1	Gerência do projeto.....	88
3.1.2	Engenharia de dados	89
3.1.3	Ciência de dados	93
3.1.4	Análise de dados	93
3.1.5	Visualização de dados.....	97
3.2	ENTENDIMENTO DO NEGÓCIO	105
3.2.1	Questionários para entendimento do uso de dados pedagógicos para avaliação de ambientes educacionais.....	106

3.2.2	Questionário para definição e avaliação de conjuntos de itens pertinentes para associação no contexto do desempenho e dos perfis socioeconômicos dos participantes do Enem 2019.....	106
3.2.3	Feedback para interpretação e avaliação dos resultados.....	107
3.3	KDD.....	107
3.3.1	Coleta de dados	107
3.3.2	Pré-processamento	107
3.3.3	Transformação	110
3.3.4	Mineração de dados	113
3.3.5	Modelo de dados para mineração	113
3.3.6	Grupos de análise - Casos de uso	115
3.3.7	Análise e assimilação de resultados.....	117
3.4	ANÁLISE DE DADOS A PARTIR DO CONJUNTO DE DADOS MODELADOS	120
3.4.1	ETL	121
3.4.2	Estatística gerais	125
4.	RESULTADOS COMPARATIVOS E ANÁLISE DIAGNÓSTICA.....	129
4.1	ESTATÍSTICAS DO ENEM.....	129
4.2	MINERAÇÃO DE DADOS	130
4.3	INTERPRETAÇÃO DOS RESULTADOS PELOS ESPECIALISTAS EM ASSISTÊNCIA SOCIAL E EDUCAÇÃO.....	134
4.4	COMPARAÇÃO ENTRE AS ANÁLISES E O SEGUNDO QUESTIONÁRIO DOS ESPECIALISTAS.....	137
5.	CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS	138
5.1	TRABALHOS FUTUROS	140

1. INTRODUÇÃO

“O Exame Nacional do Ensino Médio (Enem) foi instituído em 1998, com o objetivo de avaliar o desempenho escolar dos estudantes ao término da educação básica. O exame aperfeiçoou sua metodologia e, em 2009, passou a ser utilizado como mecanismo de acesso à educação superior, por meio do Sistema de Seleção Unificada (Sisu), do Programa Universidade para Todos (ProUni) e de convênios com instituições portuguesas. Os participantes do Enem também podem pleitear financiamento estudantil em programas do governo, como o Fundo de Financiamento Estudantil (Fies). Os resultados do Enem continuam possibilitando o desenvolvimento de estudos e indicadores educacionais.

Qualquer pessoa que já concluiu o ensino médio ou está concluindo a etapa pode fazer o Enem para acesso à educação superior. Os participantes que ainda não concluíram o ensino médio podem participar do Enem como “treineiros” e seus resultados no exame servem somente para autoavaliação de conhecimentos.” INEP.

O Enem nos dias atuais passou a ser um vetor decisivo para ingressar em universidades públicas federais e, em alguns casos, estaduais, além de também permitir ingresso em instituições de ensino privadas por meio de financiamento estudantil ou programas de bolsa universitária. Além disto, permite também o ingresso no ensino técnico subsequente em instituições conveniadas. Tendo em vista este panorama, pode-se encarar o exame como sendo muito além de um levantamento do nível educacional do ensino médio no país, mas também um veículo relevante para egressos na carreira acadêmica.

Tanto se encarado como uma forma de acesso a instituições de ensino superior e técnico quando como uma avaliação de desempenho dos estudantes no fim do ciclo da educação básica, o exame nacional do ensino médio é uma ferramenta poderosa em termos de dados e informações úteis para investigar desigualdades sociais e econômicas e disparidades geográficas, geracionais, entre outros, em termos do rendimento dos participantes nas provas e na redação. Além disto, é possível estudar a situação socioeconômica dos alunos, pois ao se inscreverem na prova, realizam o preenchimento de um questionário para levantamento destes dados.

A base de dados utilizada para aplicação das técnicas analíticas neste trabalho é a base de microdados do ENEM 2019, que pode ser extraída do portal online do INEP, que produz estes dados. A base foi baixada, extraída devidamente colocada em um repositório em nuvem de onde foi consumida para a modelagem dos dados que foi preparado para originar os

conjuntos de dados para análise estatística e para aplicação da mineração de dados a partir das técnicas do KDD e do algoritmo de descoberta de regras de associação Apriori.

O modelo arquitetado segue o projeto de um *data mart* em esquema estrela, o conjunto de dados para análise ad-hoc foi construído em um cubo OLAP e a mineração foi realizada após as transformações em SQL de uma tabela com mescla dos dados e o algoritmo foi aplicado a documentos textuais no formato CSV, extraídos a partir de consultas específicas que fracionam o conjunto de dados de acordo com os grupos de análise.

O trabalho que se segue possui dois objetivos. O primeiro é realizar um levantamento estatístico geral a partir da análise dos dados do tipo ad-hoc de caráter descritivo, para expor a estatística de todos os dados tratados segundo o objetivo analítico. O segundo é minerar os dados socioeconômicos, geográficos, pessoais, de escolaridade e do desempenho dos alunos caracterizado pela média das notas e pela nota na redação, sob a ótica do KDD em uma abordagem diagnóstica da situação do exame realizado no ano de 2019.

O trabalho aplica a arquitetura moderna de business intelligence em todo seu escopo, com escala reduzida. Foi realizada a aplicação de técnicas de engenharia de dados, para movimentação e transformação de dados entre fonte de origem e destino dos dados, conceitos e práticas de inteligência de negócios para modelar a base a ser analisada e também foram contempladas as ferramentas de análise e ciência de dados, com o algoritmo de mineração, a análise ad-hoc e a visualização de dados.

1.1. MOTIVAÇÃO E JUSTIFICATIVA

O presente estudo avaliará o impacto dos dados socioeconômicos no rendimento dos participantes do Exame Nacional do Ensino Médio, edição de 2019. Objetiva-se diagnosticar a relação entre os dados deles e sua performance nas provas que compõe o exame, que contempla todo conteúdo da base nacional comum curricular.

Este hiato acadêmico motivou o projeto, sendo importante ressaltar o vínculo pessoal que os autores possuem com o presente tema. Mediante aos seus preparatórios para as tentativas de ingressos as universidades. Sobretudo durante o período de pandemia, as disparidades econômicas e desigualdades sociais se acentuaram e o prejuízo a qualidade educacional será certamente de grande proporção.

Por isto, faz-se necessário aplicar análises descritivas e diagnósticas das bases de dados nacionais de educação que forem disponíveis para compreender quais variáveis impactam

qualitativamente no ensino de base, através de estatística e probabilidade. Com a abordagem da análise de Big Data, é possível viabilizar tecnologias de automatização para realizar estes levantamentos.

A pesquisa irá investigar por possíveis correlações de dados socioeconômicos, da realização da prova, escolares e pessoais, e através disso entender o que acarretou tais rendimentos dos participantes. Estes dados serão agrupados a média das notas das provas e a nota geral da redação, para então aferir potenciais causalidades a partir de cada um destes grupos de análise.

Por meio do apoio de especialistas, temos como objetivo traçar linhas argumentativas de interpretação das associações entre estes fatores para então diagnosticar os perfis socioeconômicos e as possíveis causas das métricas de desempenho nas áreas de conhecimento e na redação, através da associação destas variáveis com os dados supracitados.

Toda a investigação de correlações e suas possíveis causalidades, bem como descrição dos elementos do exame são organizadas dentro do contexto de um projeto de inteligência de negócios modernizado, que contemple fim a fim todas as etapas de transformação de dados através da pirâmide de conhecimento. Desta forma, pode-se obter o panorama geral dos inscritos e relacionar seu desempenho com os demais dados presentes na base.

Através do levantamento destes fatores de influência no desempenho, buscamos elucidar um caminho para avaliação da qualidade do ensino médio do país fornecendo uma ferramenta de análise orientada a dados para extração de conhecimentos e produção de sabedoria acerca do ensino médio no Brasil, que apoie em tomadas de decisão por parte de organizações da área da educação e propicie orientação para pautas de políticas públicas com foco nas vulnerabilidades educacionais do país.

Por conseguinte, esperamos contribuir com o tema apontando para a desigualdade social. E com isso expor que há vários quesitos para serem levados em consideração ao rendimento de um participante.

1.1.1. Objetivo geral

Analisar a base de micro dados do Exame Nacional do ensino médio referente ao ano de 2019 para aplicação da mineração de dados através do ciclo KDD para estruturar o conhecimento contido nela e produzir sabedoria acerca de fatores sociais e econômicos que

impactam o desempenho dos alunos participantes do exame, de modo a desvendar possíveis gargalos e problemas que afetam a qualidade do ensino médio no país.

Fornecer uma ferramenta analítica capaz de automatizar a extração de conhecimento a partir de bases de dados para produção de sabedoria para auxiliar tomada de decisões orientada a dados (Data Driven Design), através da estruturação operacional de uma engenharia de BIG DATA para ETL que viabilize o consumo de conjuntos de dados modelados das bases por parte de algoritmos de mineração de dados.

Modelar a base de dados ingerida no processo ETL e operar o algoritmo de descoberta de regras de associação APRIORI utilizando a linguagem R e o repositório CRAN, no ambiente de desenvolvimento integrado R STUDIO com uso da plataforma como serviço em nuvem Azure Databricks e criar relatórios de visualização dos data frames obtidos através da aplicação Power BI desktop e online.

Aplicar análise estatística de dados da base através da ferramenta analítica Power BI com a camada semântica de ETL em um cubo de processamento analítico on-line na plataforma como serviço em nuvem Azure Analysis Services, a fim de extrair informações da base de dados que norteiem a seleção das regras de associação com foco no escopo de pesquisa, que é socioeconômico e pedagógico.

A interpretação dos resultados e a avaliação das regras de associação mineradas foram feitas com apoio e parecer técnico dos especialistas em educação, pedagogia e assistência social, através das regras de associação obtidas que expressem as relações entre os dados avaliados.

1.1.2. Objetivos específicos

Coletar, armazenar e disponibilizar a base de dados que contém os micros dados do ENEM 2019 para modelagem.

Limpar e preparar a base de dados para que seja modelada

Entender o domínio de negócio da educação de base e levantar possíveis fatores que possam influenciar no rendimento do aluno em exames através de entrevista com especialistas.

Compreensão e visão dos dados existentes na base sobre os participantes que possam ser associados com desempenho no exame e quais as associações relevantes entre dados socioeconômicos, dados sobre o desempenho na prova e dados pessoais do participante, para

definir perfis sociais do participante, através de questionário com especialistas da educação secundária, pedagogos e assistentes sociais.

Realizar a ingestão, orquestração de movimentação de dados e o processo de ETL da base de dados para então modelá-la em um conjunto apropriado para e estruturá-la para a mineração, de modo a formatar e transformar dados em um modelo transacional para aplicação do algoritmo minerador.

Submeter o conjunto de dados modelado da base ao algoritmo de descoberta de regras de associação, que se caracteriza pelo algoritmo Apriori com consumo da biblioteca de dependência arules do ambiente R disponibilizado no repositório CRAN.

Aplicar o algoritmo AAR em diferentes cenários hipotéticos de calibração dos valores de medida distintos, de forma a testar a descoberta de regras de associação em diferentes cenários probabilísticos pré-definidos para realização de análise diagnóstica dos dados do ENEM. Produzir a organização visual dos resultados de saída da técnica de mineração da descoberta de associações.

Avaliação do modelo de dados sob a ótica da análise de dados, para obter padrões e analisar agrupamentos estatísticos dos dados presentes na base. Gerar visualização dos dados organizada e limpa para cada cenário e comparar os resultados obtidos e organizados com as informações provisionadas pelos especialistas.

Expor de forma clara e consistente aos especialistas consultados os resultados obtidos e a sua comparação em relação as prioridades que apontaram, para que elucidem e definam conclusões embasadas a partir do lhes for exposto. Coletar estas conclusões e determinar a interpretação dos resultados como última etapa do processo de descoberta de conhecimento em bancos de dados.

1.2 DIFICULDADES E LIMITAÇÕES DO TRABALHO

Nesta seção são listadas as principais dificuldades apresentadas durante todo o desenvolvimento deste trabalho.

1.1.3. Ideia inicial

Ao iniciar este trabalho, a ideia era a criação de uma Web API consumível em sistema de gestão escolar para mineração e análise de dados pedagógicos para avaliação dos rendimentos dos alunos e desempenho dos professores.

Esta ideia foi descartada, pois os dados escolares são sensíveis segundo a Lei Geral de Proteção de dados e, portanto, a base de dados de uma escola ou sua rede não poderiam ser disponibilizados, por não estarem anonimizados.

Então, decidiu-se criar uma arquitetura de business intelligence com análise e mineração de dados para ser aplicada em bases de dados públicas disponíveis no portal do INEP, que já possuem camada anônima e estão prontas para consumo e análise. Foi escolhida a base de microdados do ENEM aplicado no ano de 2019.

A ideia após a seleção da fonte e escolha da base, foi feito planejamento inicial de aplicar análise para extração de insights de toda a base e suas informações, porém o escopo do trabalho ficaria muito amplo e além do que o que poderia ser feito em uma pesquisa de monografia. As informações de inclusão e acessibilidade necessitam de uma abordagem de estudo mais profunda e especializada e, portanto, precisaria de um rigor científico maior com apoio direto de profissionais da área.

Portanto, a escolha do escopo do trabalho foi definida pautada apenas nos dados sociais e regionais a partir de informações pessoais dos estudantes, os dados de desempenho nas provas e também os dados presentes nos questionários socioeconômicos. Todos os dados de inclusão ficam como oportunidade para um trabalho futuro de análise e mineração da base sob a ótica da educação inclusiva.

1.1.4. Limitações tecnológicas

Esta seção destina-se a exposição das barreiras tecnológicas e limitações das ferramentas utilizadas.

1.1.4.1. Unificação do OLAP e do Data mining

De início, tentou-se realizar a análise de dados e a aplicação do algoritmo de mineração de dados no mesmo lugar, utilizando um modelo carregado do data lake storage para o analysis services da Azure. Este modelo seria transformado para bifurcar para as duas aplicações de tratamento de dados.

No entanto, o modelo Olap não atenderia a formatação de dados transacionais utilizadas pelo algoritmo apriori e não há nenhum conector nativo aberto para vincular uma ligação direta entre o ambiente R e o cubo OLAP do AAS. Desta forma, foi decidido utilizar um banco de dados gerenciado como serviço em nuvem da Azure, o SQL Database, para realizar a modelagem de um data mart relacional que servisse de modelo para ambas as aplicações e então realizar a bifurcação deste data mart para o conjunto de dados da mineração utilizando R e do cubo do AAS.

Posteriormente, o modelo definitivo do projeto foi desenvolvido em uma instância local de sql server 19 express e linkado com o azure data factory para ser carregado com o pipeline de dados. Os modelos de junções das tabelas para análise ad-hoc e para o data mining foram extraídos manualmente e carregados no data lake, para uso comum.

1.1.4.2. Limitação da ferramenta de Self-service BI (PowerBI)

O Power BI free possui uma limitação de funcionamento para leitura de bases de dados carregadas, pois suporta apenas 1GB de dados em um arquivo PBIX do modelo ad-hoc e realiza apenas a leitura de 1 milhão de linhas por arquivo.

Como a base, mesmo transformada, possui um tamanho de arquivo maior e uma quantidade de registros superior ao suportado, a utilização de uma camada semântica para atuar como intermediário entre a fonte de dados e o Power BI faz-se estritamente necessária.

Neste escopo entra a ferramenta de cubo para On-line Analytical Processing(OLAP), que é o analysis services em nuvem, na Azure.

1.1.5. Limitações da base de dados

A base de dados disponibilizada pelo INEP do ENEM 2019 é rica em informações dos participantes, entretanto nós detectamos a falta de algumas informações que julgamos ser importante e há informações que poderiam ser substituídas, devido estarem obsoletas e terem perdido o seu valor, tendo em vista que vivemos em uma sociedade que está em constante desenvolvimento.

Podemos destacar que a coleta dos dados de estado e cidade de nascimento dos inscritos só são coletadas quando o inscrito é brasileiro, mas quando os inscritos são brasileiros

naturalizados, brasileiros natos nascido no exterior ou estrangeiros, não há coleta desses dados, além disso não a origem do seu país, isso poderia agregar valores nas análises dos dados.

Entendemos que as colunas de cidade e estado de residência dos inscritos, são itens obrigatório para o cadastro dos participantes, nesse caso, isso inviabiliza o acesso ao exame para pessoas sem residência e para pessoas que vivem em outro país, mas cruzam a fronteira para realizá-la, sem necessariamente estarem vivendo no Brasil.

Outra informação que podemos contestar é na coluna de sexo, onde há a opção de masculino e feminino, todavia não há opção de não declaração do gênero, como há a opção de não declaração de cor e raça.

Um dado crucial que o INEP possui e poderia adicionar a base de dados que seria de grande valor seria informação sobre os inscritos que tiveram isenção, pois isso também auxiliaria na montagem e entendimento dos relatórios, baseado nesse tipo de informação.

Não há uma especificação de onde está localizado a residência dos inscritos como está descrito por exemplo as escolas do ensino médio, como rural e urbana, pois isso também facilitaria o entendimento sobre a localidade vivida dos inscritos, pois atualmente podemos ter uma ideia sobre os inscritos que estão cursando o ensino médio, devido a localidade de sua escola, mas não dos demais.

É visto que falta a informação de quais inscritos que estão privados da liberdade, que realizam o ENEM PPL (Exame Nacional do Ensino Médio para Pessoas Privadas de Liberdade) fizeram a inscrição para o exame.

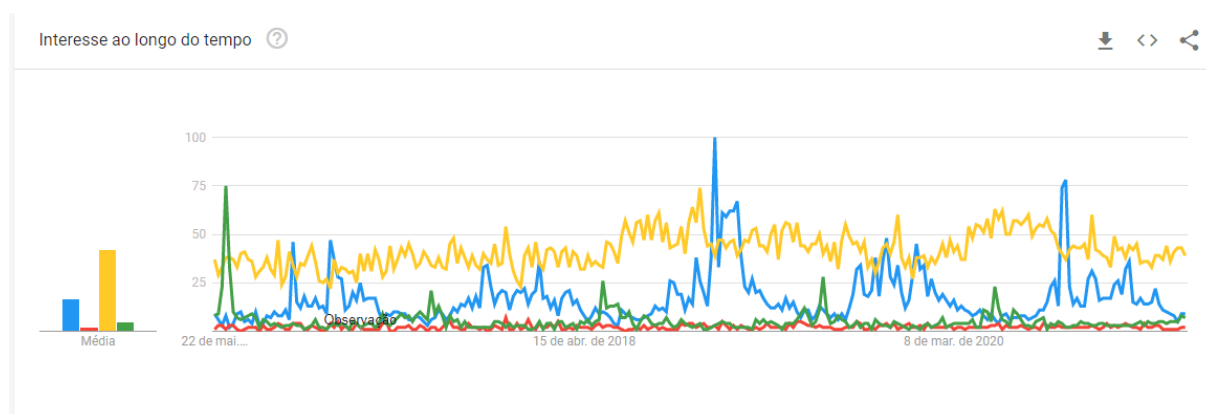
Uma informação que poderia ser acrescentada a base de dados de informações correspondente aos pais, seria o grau de relacionamento atual entre os pais, pois segundo Moreira (2010) baseado em entrevistas com professores e analistas podemos chegar à conclusão de que a separação dos pais afeta na aprendizagem do aluno. Ainda falando sobre as informações dos pais ou responsáveis, nas informações sobre a ocupação dos responsáveis não há informações sobre beneficiários BPC, aposentados, pensionista, desempregado ou desalento, nas notas de rodapé a respeito do questionário informa que se o responsável não estiver trabalhando deve ser marcado a última profissão, entretanto entendemos que isso pode causar um desequilíbrio na análise dos dados.

No questionário há questões sobre a estrutura da residência, mas sobre a quantidade e existência dos quartos e banheiros, embora esses itens sejam relevantes, não há questões sobre luz elétrica, água encanada, esgoto e suas variações, além disso poderiam ser questionados se

utilizam gás encanado, botijão ou até fogão a lenha. Sobre a residência não possuímos informações sobre o tipo de residência, se é alugada, própria, cedida, etc.

Podemos destacar que no questionário de informações socioeconômica que abrange os eletrodomésticos não constam as informações sobre a possibilidade de os inscritos possuir um fogão, ar-condicionado, aquecedor e ferro de passar. Utilizamos o Google Trends para fazer uma análise de interesse sobre esses itens e podemos verificar na figura 1 a distribuição do interesse ao longo do tempo, foram utilizados alguns filtros para podermos ser mais específicos, como o da região que está configurado para podermos ver somente o Brasil, o período é nos últimos 5 anos e o local de pesquisa sendo o Google Shopping, já que estamos nos baseando na busca pela aquisição desses itens.

Figura 1 - Gráfico de linha contendo análise de tempo de interesse sobre fogão, ar-condicionado, aquecedor e ferro de passar.

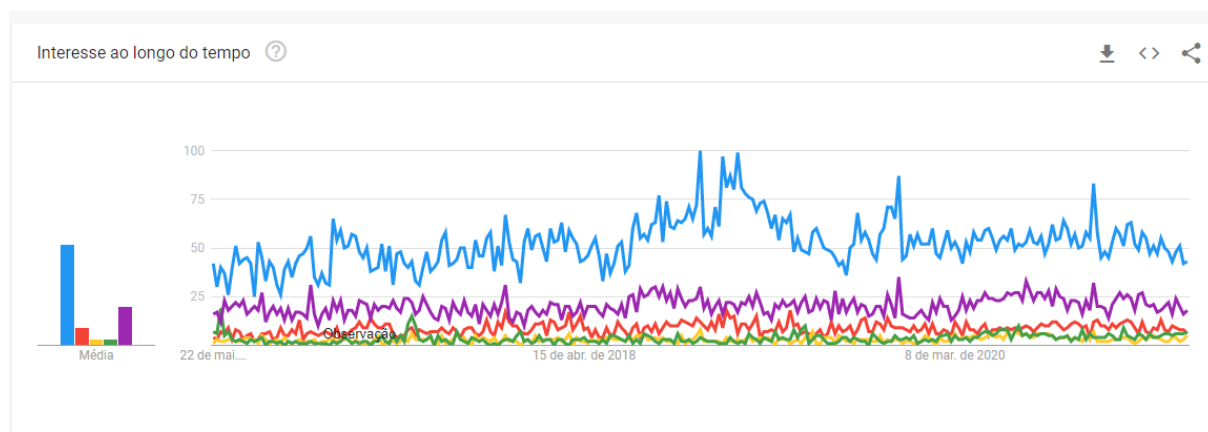


Fonte: Print da pesquisa do Google Trends

O azul corresponde ao ar-condicionado que contém 17 pontos de interesse, o vermelho ao ferro de passar que contém 2 pontos de interesse, o amarelo ao fogão que contém 42 pontos de interesse e o verde ao aquecedor que contém 45 pontos de interesse.

Realizamos mais algumas pesquisas dentro do Google Trend utilizando os mesmos parâmetros supracitados, para os itens que estão na base de dados do ENEM.

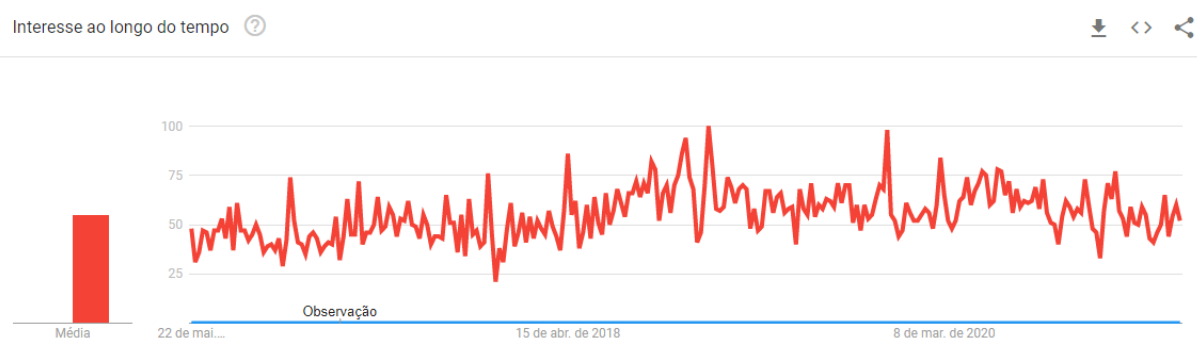
Figura 2 - Gráfico de linha contendo análise de tempo de interesse sobre geladeira, freezer, máquina de lavar louça, secadora de roupas e forno de micro-ondas.



Fonte: Elaborado pelo autor através do Google Trends (2021)

Na figura 2, o azul corresponde a geladeira que contém 52 pontos de interesse, o vermelho corresponde ao freezer que contém 9 pontos de interesse, o amarelo corresponde a máquina de lavar louça que contém 3 pontos de interesse, o verde corresponde a secadora de roupas que contém 3 pontos de interesse e o lilás corresponde ao forno micro-ondas que contém 20 pontos de interesse.

Figura 3 - Gráfico de linha contendo análise de tempo de interesse sobre o aspirado de pó e sobre a lavadora de roupas.



Fonte: Elaborado pelo autor através Google Trends (2021)

Na figura 3 o azul corresponde o aspirador de pó que não contém nenhum ponto de interesse e o vermelho corresponde a lavadora de roupas que contém 55 pontos de interesse

Baseando-se nessa pesquisa podemos expressar esses pontos através da tabela 1, que mostra de forma classificada, do que possui mais interesse para o que possui menos.

Tabela 1 - Classificação dos eletrodomésticos pontuados através do Google Trends

Produto	Pontos de interesse	Base de dados do ENEM
Lavadora de roupas	55 pontos	Está na base de dados
Geladeira	52 pontos	Está na base de dados
Fogão	42 pontos	Não está na base de dados
Forno micro-ondas	20 pontos	Está na base de dados
Ar-condicionado	17 pontos	Não está na base de dados
Freezer	9 pontos	Está na base de dados
Aquecedor	5 pontos	Não está na base de dados
Secadora de roupas	3 pontos	Está na base de dados
Máquina de lavar louça	3 pontos	Está na base de dados
Ferro de passar	2 pontos	Não está na base de dados
Aspirador de pó	0 pontos	Está na base de dados

Fonte: Elaborado pelo autor (2021)

Conforme destacamos na tabela acima, há itens que não estão na lista de eletrodomésticos, mas são também tão procurados e adquiridos, quanto os outros que já estão no questionário. Acreditamos que não seja necessário remover nenhum, mas sim, acrescentar.

Ainda há na base de dados a questão que implica na quantidade e existência de televisão em cores, entretanto não questiona sobre a existência ou quantidade televisões com acesso à internet ou dispositivos que possibilitam a televisão ter acesso a internet, isso por ser mais um recurso que tem acesso à internet e de certa forma pode favorecer ao inscrito. Da mesma forma também não é conferido se na residência do inscrito há assinaturas de serviços *stream*, entretanto nas especificações multimídia tem a opção de DVD, mesmo sendo obsoleto, e não há a opção de consoles de jogos, pois é uma central multimídia. Averiguamos que também não há a especificação do tablet ou aparelhos de leitura de livro digital, que podem auxiliar no

estudo dos inscritos. Além dessas informações não constam a diferença os tipos de computadores em desktop e portáteis, justamente pela facilidade de locomoção.

Já sobre a internet, ela não é especificada como fixa ou móvel.

A única inconsistência que encontramos na base de dados foi na coluna STATUS da Redação onde falta um número, como pode ser visto na figura 4.

Figura 4 - Informações da coluna de status de redação.

Situação da redação do participante	1	Sem problemas
	2	Anulada
	3	Cópia Texto Motivador
	4	Em Branco
	6	Fuga ao tema
	7	Não atendimento ao tipo textual
	8	Texto insuficiente
	9	Parte desconectada

Fonte: Print do Dicionário de Dados dos Micro dados no ENEM de 2019

Podemos constatar a ausência do número 5.

1.1.6. Custos de infraestrutura

O Azure tem vários tipos de assinaturas. A assinatura é um contrato assinado com a Microsoft para usar uma ou mais plataformas ou serviços em nuvem da Microsoft. A taxa é baseada na taxa de licença de cada usuário ou com base no consumo de recursos da nuvem.

Nos serviços de nuvem baseados em Azure PaaS, as licenças de software foram integradas ao preço do serviço.

Para máquinas virtuais baseadas em Azure IaaS, pode ser necessário ter outras licenças para usar o software ou aplicativos instalados na imagem da máquina virtual. Algumas imagens de máquina virtual já possuem uma versão licenciada do software instalada e a taxa está incluída no custo por minuto do servidor.

O Azure tem uma assinatura de estudante gratuita, que permite aos usuários começar a usar o Azure com um crédito de US \$ 100. Além de escolher serviços gratuitos e não exigir um cartão de crédito ao se registrar, o crédito também pode ser usado nos primeiros 12 meses.

Esta assinatura está disponível apenas para alunos que atendam aos seguintes requisitos. É necessário confirmar que o usuário tem no mínimo 18 anos, e que foi admitido em um curso de graduação de 2 a 4 anos por uma instituição de ensino reconhecida, e que o usuário é

estudante em período integral. Os usuários devem usar o endereço de e-mail de sua organização para verificar seu status acadêmico.

Esta assinatura fornecerá acesso a certos benefícios de download de software, cujo propósito expresso é apoiar a educação, pesquisa não comercial ou esforços para criar, desenvolver, testar e demonstrar aplicativos de software para os fins mencionados supracitados.

Utilizamos três assinaturas, duas assinaturas de estudante e uma assinatura grátis para usuário comum.

Figura 5 - Gastos da assinatura gratuita para usuário comum



Fonte: Elaborado pelo autor através do da Azure

Figura 6 - Gastos do usuário 1 da assinatura para estudante

ServiceName	Cost
Advanced Data Security	\$0,00
Azure Analysis Services	\$0,07
Azure Data Factory v2	\$17,93
Bandwidth	\$0,00
SQL Database	\$235,65
Storage	\$0,00
Virtual Network	\$0,01
Total	\$253,66

Fonte: Print do Power BI

Figura 7 - Gastos do usuário 2 na assinatura para estudante

ServiceName	Cost
Azure Analysis Services	\$52,80
Azure Bastion	\$3,42
Azure Data Factory v2	\$2,35
Bandwidth	\$6,44
Data Lake Store	\$0,24
Storage	\$11,57
Virtual Machines	\$2,21
Virtual Network	\$6,89
Total	\$85,93

Fonte: Print do Power BI

Nas figuras 5, 6 e 7, podemos ver os gastos de cada assinatura e de cada serviço utilizado.

Nota-se que assinatura de estudante do usuário 1 excedeu os créditos utilizados, isso devido ao serviço de SQL, entretanto a conta sofreu um bloqueio de escrita, ela pode ser acessada, mas somente para consulta.

1.1.6.1. Azure Data Lake

O custo total do Azure Data Lake Store Gen1 depende da quantidade de armazenamento, como o número e o tamanho das transferências e transações de dados de saída.

O armazenamento é dividido em pacotes pay-per-use e pacotes de compromisso mensal.

Tabela 2 - Pós Pago

Uso	Preço/mês
Primeiros 100 TB	R\$0,1911 por GB
Próximos 100 TB a 1.000 TB	R\$0,1862 por GB
Próximos 1.000 TB a 5.000 TB	R\$0,1813 por GB
Mais de 5.000 TB	Valor gerado através de cotação da Microsoft

Fonte: Azure Microsoft

Comparado ao preço pago com base no uso, o pacote de compromisso mensal tem um desconto significativo (até 33%). O restante do armazenamento incluído será cobrado a R \$ 0,1911 / GB / mês.

Tabela 3 - Pacotes de compromisso mensal

Capacidade comprometida	Preço/mês	Economia no pago conforme o uso
1 TB	R\$171,462	12%
10 TB	R\$1.567,648	20%
100 TB	R\$14.206,810	27%
500 TB	R\$65.155,37	32%
1.000 TB	R\$127.371,400	33%
Mais de 1.000 TB	Valor gerado através de cotação da Microsoft	

Fonte: Azure Microsoft

O próximo valor se aplica às transações executadas com os dados. Tanto o pacote de uso pago quanto o pacote de compromisso mensal cobram a mesma taxa de transação.

Tabela 4 - Transição de dados

Uso	Preço
Operações de gravação (por 10.000)	R\$0,245
Operações de leitura (por 10.000)	R\$0,0196
Operações de exclusão	Gratuito
Limite de tamanho de transação	Sem limite

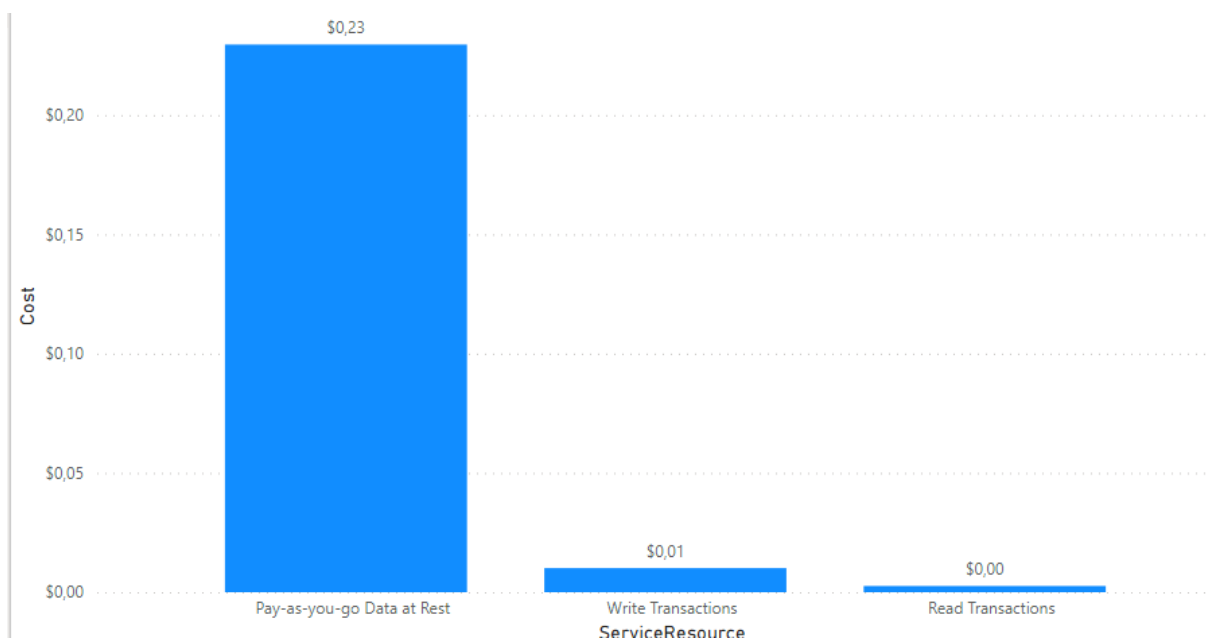
Fonte: Azure Microsoft

Os acordos de suporte e nível de serviço são gratuitos para o gerenciamento de cobranças e assinaturas, mas para planos de suporte flexível, o preço inicial é de R \$ 142.069 / mês. Exceto para serviços de visualização, a disponibilidade de garantia é de 99,9% ou mais.

Qualquer operação no Azure Data Lake Store Gen1 é cobrada como uma única transação. Isso inclui operações HTTP e operações concluídas por trabalhos do Azure Data Lake Analytics.

Nossos custos com o Data Lake foram relativamente baixo como mostra a figura 8, podemos ver o gasto pela utilização dos recursos.

Figura 8 - Gasto dos recursos do Data Lake do usuário 2



Fonte: Print Power BI

A nossa utilização do Data Lake foi para armazenar os dados originais e os dados transformados. Através do Data Lake, nós conectamos o Analysis Services e outros serviços da Azure e Microsoft. A utilização foi por 59 dias.

1.1.6.2. Azure SQL Data base

O banco de dados SQL do Azure faz parte da série de serviços de banco de dados SQL do Azure SQL. É um serviço de banco de dados inteligente e escalonável criado para a nuvem. Ele tem funções de plataforma AI e pode manter o desempenho e a durabilidade ideais. Por meio da computação sem servidor e da expansão automática de recursos de armazenamento

ultra grandes, os custos podem ser otimizados sem se preocupar com o gerenciamento de recursos.

Buscando flexibilidade, controle e transparência no uso de um único recurso, o modelo de compra baseado em vCore será melhor. Esse modelo permite o dimensionamento de computação, memória e armazenamento de acordo com as necessidades da carga de trabalho e é uma maneira fácil de mover as necessidades de cargas de trabalho locais para a nuvem.

A camada de computação sem servidor do SQL Server otimiza o desempenho de preço e simplifica o gerenciamento de desempenho de um único banco de dados com uso intermitente e imprevisível, dimensionando cálculos automaticamente e cobrando por cálculos usados por segundo.

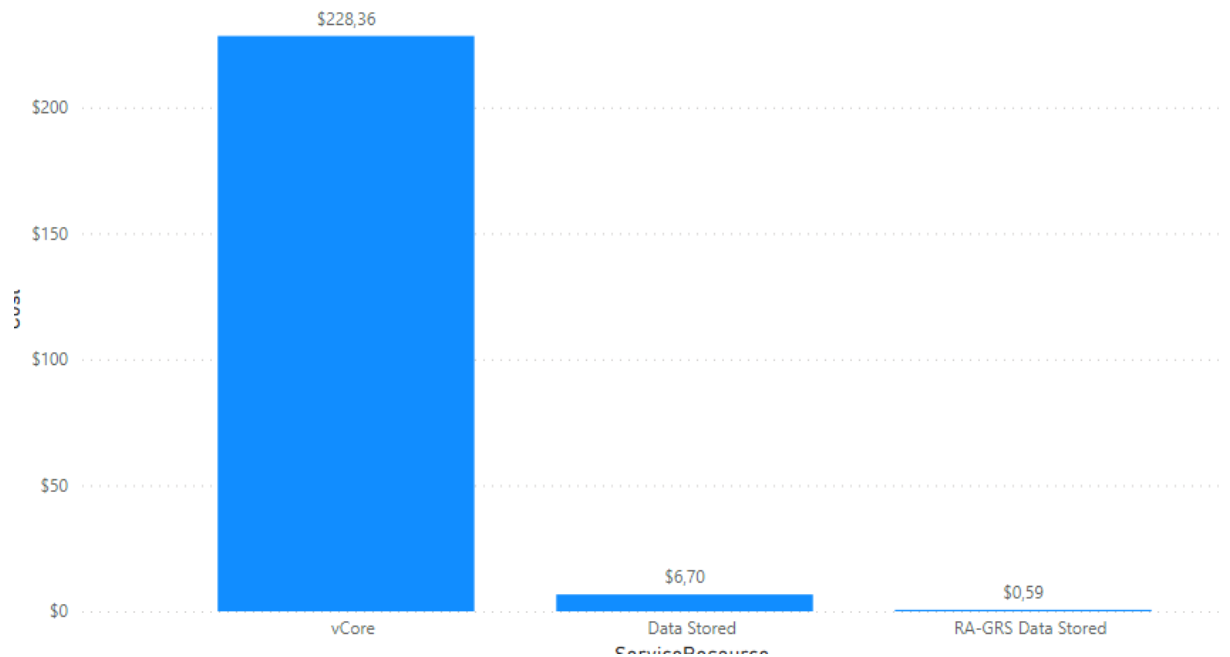
Como listado na tabela de custos de infraestrutura, a manutenção da estrutura de um banco de dados como serviço gerenciável e sem servidor(serverless) como o Azure SQL Database é bastante alta. Isto inviabilizou o uso da tecnologia em nuvem para a produção final do trabalho, sendo então utilizado apenas como uma ferramenta de teste para homologação do modelo.

O desenvolvimento do projeto final maturado do modelo de banco de dados no esquema de relational data mart foi inteiramente desenvolvido em uma instância de banco de dados Microsoft SQL Server Express 2019 em uma máquina local, que é o mesmo banco utilizado como tecnologia concreta no back-end do PaaS Azure SQL DB, que abstrai o MSSQL Server em sua versão mais estável.

Todo o processo de pipeline para cópia de dados com uso da ferramenta de ingestão e orquestração de movimentação e transporte de dados PaaS da Microsoft Azure Data Factory tinha sido realizada com o banco de dados em nuvem e foi efetuado com sucesso tendo como destino o banco de dados local MSSQL, através da configuração de um runtime de integração auto-hospedado para tornar a máquina um servidor ADF.

Assim, o banco na máquina local foi o destino da ingestão de dados no ETL e foi modelado de acordo com a necessidade do presente trabalho.

Figura 9 - Gasto dos recursos do SQL do Azure do usuário 1



Fonte: Print do Power BI

1.1.6.3. Analysis Services

O Azure Analysis Services fornece recursos de modelagem semântica de BI de nível empresarial e tem as vantagens de escala, flexibilidade e gerenciamento fornecidas pela nuvem. O Azure Analysis Services pode ajudá-lo a transformar dados complexos em informações acionáveis. O Azure Analysis Services inclui um mecanismo de análise confiável do Microsoft SQL Server Analysis Services.

O custo total do Azure Analysis Services depende da camada e instância selecionadas.

O Azure Analysis Services está disponível nas camadas de Desenvolvedor, Básica e Standard.

As principais diferenças entre os recursos de cada camada estão nos atributos de perspectivas, variedade de partições e modo de armazenamento DirectQuery

Em cada camada, o preço de uma instância varia de acordo com o poder de processamento, QPU e tamanho da memória.

A camada básica é para uso geral e é recomendada para soluções de produção que têm um modelo tabular pequeno, simultaneidade de usuário limitada e requisitos simples de atualização de dados.

Tabela 5 - Preço Analysis Services

Instância	QPU	Memória (GB)	Preço
B1	40	10	R\$1.537,765/mês

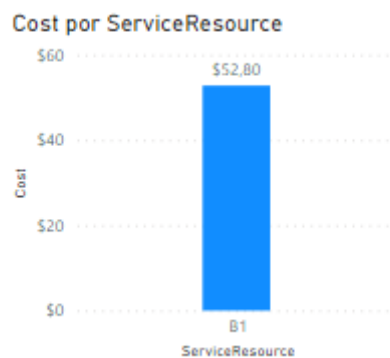
Fonte: Azure Microsoft

Aplicam-se taxas de transferência de dados padrão. A estimativa de preço mensal é baseada em 730 horas por mês.

O Analysis Services do Azure cobra uma taxa previsível por hora com base na camada de serviço e no nível de desempenho de uma instância única. O uso real é calculado para o segundo e cobrado por hora.

Nosso período de uso do Analysis Service foi de 26 dias, utilizamos com pausa, só o tirávamos da pausa para a utilização, assim que o uso não era mais necessário, nós desligávamos para economizar os créditos.

Figura 10 - Gastos do Analysis Services, Recuso B1



Fonte: Elaborado pelo autor

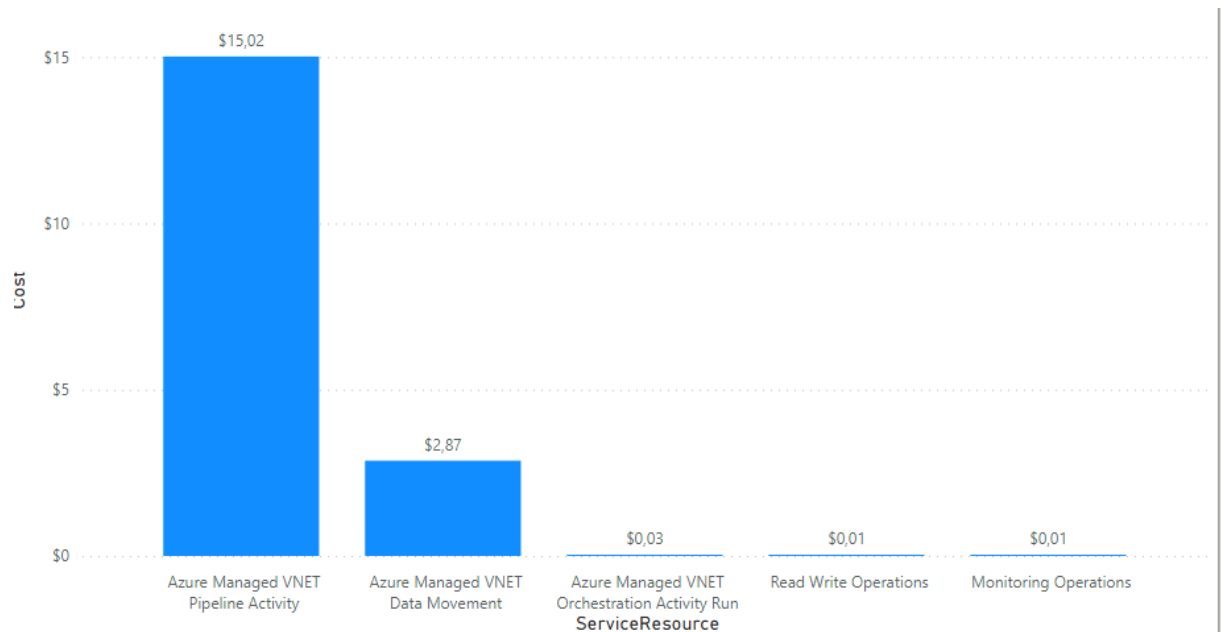
1.1.6.4. Data factory

O Azure Data Factory é um serviço elástico de integração de dados sem servidor projetado para escala de nuvem. Isso significa que não há necessidade de planejar cálculos de tamanho fixo para cargas de pico. Em vez disso, ele especifica a quantidade de recursos alocados para cada operação sob demanda, o que permite o design de processos ETL de uma

forma mais escalonável. Além disso, o ADF é cobrado de acordo com o plano de consumo, o que significa que você paga apenas pelas mercadorias que utiliza.

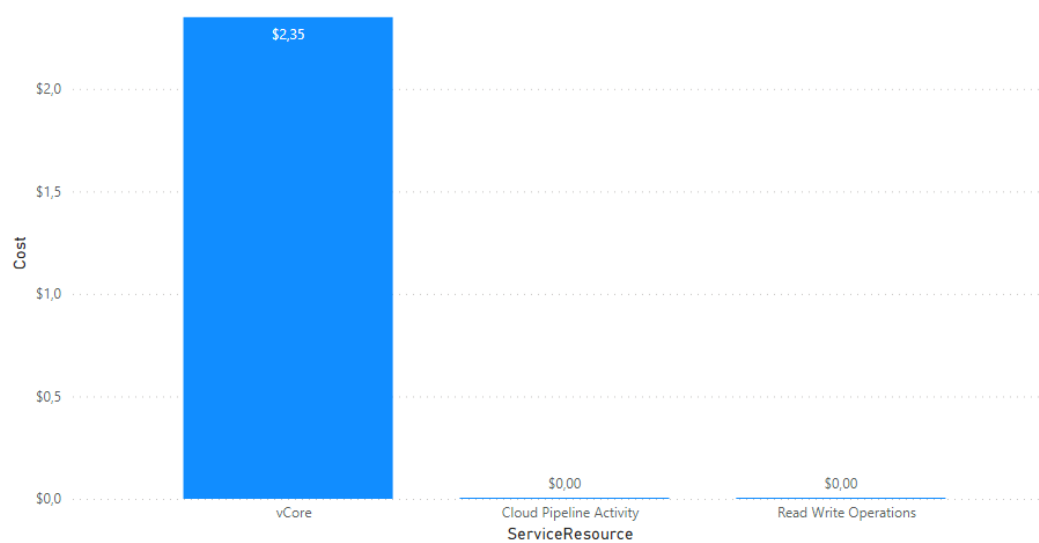
O Data Factory foi utilizado por 9 dias na assinatura para estudantes do usuário 1 e por 2 dias na assinatura de estudante do usuário 2.

Figura 11 - Gastos dos recursos do Data Factory da assinatura de estudante do usuário 1



Fonte: Elaborado pelo próprio autor (2021)

Figura 12 - Gastos por recursos



Fonte: Elaborado pelo auto (2021)

1.3 CONTEXTUALIZAÇÃO DO PROBLEMA

O exame nacional do ensino médio é o principal meio de ingresso para as universidades públicas no país, federais e estaduais, mas além disto, é uma forma de medir o nível de qualidade do ensino no país, por integrar toda a base nacional comum curricular no conteúdo programático do exame. Como trata-se de uma prova extensível em escala nacional e geracional, disponível para alunos a partir do 2º ano do ensino médio e para pessoas que já o concluíram, é uma fonte de análise com potencial para aferir o nível educacional do país de forma abrangente em dimensões geográficas, históricas, sociais e econômicas.

Ao analisar os dados dos participantes presentes na base de micro dados do instituto nacional de pesquisas educacionais Anísio Teixeira (INEP), pode-se extrair insights e produzir conhecimentos e sabedoria através deles na direção da detecção das variáveis que estão diretamente relacionadas com níveis de desempenho variados e traçar indicadores socioeconômicos e regionais que apoiem na avaliação do rendimento no ENEM e, por consequência, apoie na identificação de impactos educacionais no ensino médio do país.

1.1.7. Escopo do trabalho

O trabalho que se segue é delimitado a análise estatística e mineração de dados socioeconômicos e de desempenho educacional na base de micro dados do ENEM fornecida pelo INEP, que se encontra disponível publicamente para ser baixada em seu portal online.

Para isto, utiliza-se tecnologias de manipulação e armazenamento dos dados, ferramentas operacionais de apoio e execução nos processos de orquestração e tratamento do conjunto de dados e metodologia analítica para aplicação da ciência de dados, que se caracteriza pelo ciclo de descoberta de conhecimento em bases de dados (KDD).

O objetivo final é desenvolver uma série de procedimentos para gestão e modelagem da base de dados baseada no foco estabelecido para a mineração, que é o estudo de informações socioeconômicas e pedagógicas, e aplicar o algoritmo minerador designado, que é o de descoberta de regras de associação Apriori (AAR).

A intenção do trabalho é, especificamente, avaliar o impacto de desigualdades sociais e econômicas no rendimento dos participantes do exame de modo a levantar fatores de influência em seu desempenho no ensino médio, através de ciência de dados e utilizar a análise dos dados preparados para obter estatísticas que demonstrem o panorama geral destes participantes, segundo estas métricas analisadas (fatos sociais, condições econômicas e desempenho na prova e informações pessoais).

O estudo dos dados sobre inclusão não foi aprofundado, sendo resumido apenas a avaliação da existência ou não de necessidade especial ou específica, pois não é o foco do trabalho atual e, portanto, coloca-se como uma oportunidade de trabalho futuro.

1.4 STACK DE TECNOLOGIAS E FERRAMENTAS ESCOLHIDAS

Foram utilizadas as seguintes ferramentas e tecnologias para a aplicação dos processos executados neste trabalho:

Armazenamento dos dados: Azure Data Lake Storage Gen1 (Tecnologia) e Azure Data Explorer (Ferramenta)

Ingestão e ETL de dados: Azure Data Factory (ferramenta PaaS) e Mapping data flows (tecnologia)

Modelagem de Data Warehouse e Banco de Dados: Azure SQL Database (Ferramenta PaaS) SQL Server 2019 Express (ferramenta On-Premises), SGBD Azure Data Studio (Ferramenta On-Premises) e SQL Server Management Studio (ferramenta On-Premises)

Modelagem de Cubo OLAP e Camada Semântica de BI: Azure Analysis Services (Ferramenta PaaS) e IDE Visual Studio Community (Ferramenta On-Premises)

Business Intelligence e Análise de dados: Power BI (ferramenta) e DAX Language (Tecnologia)

Ciência e mineração de dados: Azure Databricks (Ferramenta PaaS) R Language (tecnologia) e IDE R Studio (ferramenta On-Premises)

Modelo de formatação de arquivos brutos: Comma-Separated Values ou CSV (tecnologia)

1.5 ANONIMIZAÇÃO DE DADOS SENSÍVEIS E LGPD

O INEP utiliza anonimização dos dados para os microdados do ENEM para que não seja possível identificar as pessoas que estão na base de dados, conforme se depreende art. 5 inciso XI da Lei Geral de Proteção de Dados Pessoais, *in verbis*:

Art. 5º. Para os fins desta Lei, considera-se:

XI - anonimização: utilização de meios técnicos razoáveis e disponíveis no momento do tratamento, por meio dos quais um dado perde a possibilidade de associação, direta ou indireta, a um indivíduo;

Na nota de rodapé do dicionário de micro dados há uma nota que informa assim,

Referente ao Enem 2019, trata-se de uma máscara e não o seu número de inscrição original no Enem. O mesmo NU_INSCRICAO para anos diferentes não identifica o mesmo participante no exame, não permite o acesso aos dados cadastrais como nome, endereço, RG etc, nem identifica o mesmo participante em microdados de pesquisas diferentes (INEP, 2020).

Com essas informações anonimizadas o INEP fica dentro das normas da LGPD, assim evitando qualquer tipo de correlação dos dados da base com algum indivíduo.

2. BIG DATA, ANÁLISE DE DADOS E O ENEM

2.1 INEP

O Instituto Nacional de Pesquisas Educacionais Anísio Teixeira, mas conhecido apenas como "INEP", foi criado há mais de 80 anos e é na verdade o braço direito do Ministério da Educação (MEC). Isso porque se trata de uma instituição que se responsabiliza pelas ações educativas realizadas no Brasil para o desenvolvimento do país.

O INEP é responsável pela organização e aplicação do ENEM, mas o Instituto também desempenha muitas outras funções. Uma delas é o desenvolvimento do Exame Nacional de Desempenho dos Estudantes (ENADE), aplicável aos alunos dos anos finais da graduação e é um dos indicadores utilizados nas avaliações atribuídas a cada professor pelo MEC.

Outros indicadores educacionais também são de responsabilidade do INEP:

- Conceito Preliminar de Curso;
- Sistema Nacional da Educação Básica;
- Índice de Desenvolvimento da Educação Básica;
- Sistema Nacional de Avaliação da Educação Superior;
- Avaliação externa das faculdades.

Por meio de pesquisas e avaliação do sistema educacional brasileiro, o INEP investiga ações de melhoria da educação básica e superior.

Por ser um instituto de pesquisa, é responsável pela realização de censos educacionais e avaliações diversas.

Portanto o objetivo do instituto é fazer com que os planos e projetos de governo na área da educação sejam formulados e implementados, garantindo assim o desenvolvimento educacional, social e até econômico do país. Por meio de avaliação e pesquisa realizadas pelos INEP, é possível encontrar carências e necessidades na educação brasileira e atuar contra essas lacunas.

2.2 EXAME NACIONAL DO ENSINO MÉDIO

O Exame Nacional do Ensino Médio (ENEM) consiste em um exame anual que é aplicado em dois dias, geralmente entre o fim de outubro e o início de novembro.

Atualmente, a prova do ENEM é realizada em dois domingos consecutivos, com duas aplicações distintas, uma para a versão impressa e outra para a versão digital.

O Enem possui 180 questões objetivas e uma redação. As questões são subdivididas entre quatro áreas do conhecimento que abrangem o conteúdo curricular dos três anos do ensino médio.:

- Linguagens, Códigos e suas Tecnologias: Língua Portuguesa, Literatura, Artes, Educação Física, Tecnologias da Informação e Comunicação e Língua Estrangeira (Inglês ou Espanhol)
- Matemática e suas Tecnologias: Matemática
- Ciências Humanas e suas Tecnologias: História, Geografia, Filosofia e Sociologia
- Ciências da Natureza e suas Tecnologias: Química, Física e Biologia

No primeiro dia, os alunos devem responder a perguntas sobre linguagens, Códigos e suas Tecnologias e também fazer a redação, possuindo um tempo total de 5 horas e 30 minutos.

No segundo dia, é a vez do candidato testar os conhecimentos de ciências naturais e de conteúdos relacionados à tecnologia relacionados à matemática e sua tecnologia. Nesta fase, o tempo de prova é ainda menor: 5 horas.

O ENEM foi criado em 1998 para avaliar o desempenho de egressos do ensino médio. Desde 2004, o exame é utilizado como ferramenta de ingresso em instituições de ensino superior e foi incluído no Sistema de Seleção Unificada (SISU) em 2010. É reconhecido como o maior e mais completo exame educacional do Brasil.

O ENEM também é o método utilizado pelos alunos para a obtenção de recursos do governo federal: Fundo de Financiamento Estudantil (FIES) e Programa de Financiamento Estudantil (P-FIES). O exame também inclui alunos que realizaram cursos técnicos por meio do Sistema de Seleção Unificada da Educação Profissional e Tecnológica (SISUTEC). Milhares de alunos participam das avaliações todos os anos.

É opcional a participação no ENEM, mas embora não seja obrigatório, é altamente recomendável realização do exame, pois através dele é possível:

- Ingressar nas universidades por meio de qualquer programa do governo federal para expandir as oportunidades de ensino superior.
- Para pessoas que não concluíram seus estudos a tempo e não frequentaram uma escola regular, podem passar no exame para obter um certificado do ensino médio.
- Serve também para melhorar a pontuação no vestibular que aceita o ENEM como complemento de sua pontuação.

- Elimina a necessidade de fazer exames de admissão para universidades privadas.

A importância desse exame para o estudante é que o ENEM é um exame que explora diversas áreas do conhecimento e aplica o conceito de interdisciplinaridade na elaboração da prova. Assim, os estudantes que realizam o exame têm o desafio de demonstrar sua capacidade de raciocínio lógico e interpretação.

Além do mais, ele prevê que o candidato seja capaz de desenvolver uma redação com elementos de dissertação, desenvolvendo seus argumentos e defendendo seu ponto de vista. Quando obtém uma boa nota no ENEM, o estudante está credenciado para ingressar em universidades públicas e privadas concorridas. Inclusive, o estudante pode conseguir acesso a algumas universidades no exterior.

O público-alvo do ENEM são todos os estudantes que estão no ensino médio ou que já concluíram o ensino básico. No caso dos estudantes do 1º e do 2º ano do ensino médio, eles participam do exame como treineiros, ou seja, não poderão usar as notas para ingressar no ensino superior.

E ainda, os alunos de supletivo e os aprovados no Exame Nacional para Certificação de Competências de Jovens e Adultos (ENCCEJA) também podem fazer o ENEM.

2.2.1 Acessibilidade e Inclusão social

A Política de Acessibilidade e Inclusão do Instituto Nacional de Educação Anísio Teixeira (INEP) organiza o Exame Nacional do Ensino Médio (ENEM), que garante auxílio profissional a determinados participantes no dia do exame. Pessoas com baixa visão, cegueira, deficiência física, deficiência auditiva, deficiência intelectual (mental), surdez, surdez-mude, dislexia, autismo, mulheres grávidas, mulheres que estejam amamentando, idosos, estudantes hospitalizados e / ou outros podem solicitar ajuda em situações específicas. São recursos de acessibilidade que podem tornar o espaço ou tempo do teste mais adequado. Salas de fácil acesso, tempo extra, aparelhos auditivos, tradução em linguagem de sinais, tratamento por nome social e acompanhantes são alguns desses exemplos.

Salienta-se que o ENEM não é um exame gratuito e, ainda que possua um valor inferior comparado com os principais vestibulares do país, há diversas pessoas que não conseguem arcar com os custos de inscrição. O Ministério da Educação (MEC) instituiu parâmetros para a gratuidade da taxa para determinados grupos, fundamentados na renda e na escolaridade dos participantes.

A isenção pode ser requerida pelos determinados grupo de participantes:

- Estudantes que estão no terceiro ano do ensino médio em escolas públicas;
- Participantes que se enquadrem na Lei Federal nº 12.799/2013;
- Inscritos no Cadastro Único para Programas Sociais do Governo Federal:

2.3 BIG DATA

Big data significa grandes dados ou mega dados. Em outras palavras, big data é um esforço para extrair informações de grandes quantidades de dados. Mas não é necessário apenas extrair o conteúdo, mas também dar-lhe sentido, e usar isso para orientar estratégias e ações (BLOG DO EAD UCS, 2020).

Sendo assim, big data é um conceito que descreve a grande quantidade de dados estruturados e não estruturados gerados a cada segundo (NASCIMENTO, 2017).

As diferenças de big data estão intimamente relacionadas à possibilidade e oportunidade de cruzar esses dados de diferentes fontes para obter insights rápidos e valiosos.

A essência do conceito é criar valor para a empresa. Quanto mais dados há, mais trabalho de processamento para gerar as informações. Portanto, a rapidez na obtenção de informações faz parte do sucesso que o Big Data pode proporcionar para os negócios.

O conceito de big data considera três pilares principais eles são, volumes, velocidade e variedade.

Para transformar os dados em informações inteligentes, é necessário seguir um ciclo que consiste em três ações principais, que são integrar, gerenciar e analisar.

Talvez o maior desafio seja implementar a reestruturação necessária. Isso porque, para executar o big data com excelente desempenho, é necessária uma grande infraestrutura técnica de suporte ao processamento dos dados. Nesse sentido, devido ao trabalho árduo, os profissionais muitas vezes encontram algumas resistências. Afinal, todos devem cooperar para revisar o processo antigo e criar por sua vez. Outro entrave que pode dificultar a implantação do big data é a falta de mão de obra qualificada. Por se tratar de uma profissão relativamente nova, é difícil encontrar especialistas neste campo. Por outro lado, é uma oportunidade promissora para quem deseja construir uma carreira de sucesso. (BLOG DO EAD UCS, 2020)

2.3.1 Big Data Analytics

A análise de big data é onde as técnicas de análise avançadas operam em grandes conjuntos de dados. Portanto, a análise de big data envolve duas coisas - big data e análise, além de como os dois juntos criam uma das tendências mais profundas da inteligência de negócios de hoje.

O processo de análise de dados criou o princípio para isso. Tudo isso devido a uma análise muito cuidadosa e precisa. Todos esses modos são projetados para filtrar e fornecer todas as informações úteis. Por trás disso, existe um ciclo que vai desde a extração dos dados, até a organização, processamento e análise dos dados. Quando se trata de análise de big data, podemos dizer que sua aplicação ocorre de três formas principais, sendo elas social data, enterprise data e personal data.

Dessa forma, os principais benefícios que a análise de big data pode gerar são a análise competitiva, a determinação de padrões precisos, a redução de custos, a tomada de decisões e o desenvolvimento de produtos e serviços.

2.4 Engenharia de dados e Big Data

A engenharia de dados inclui o desenvolvimento e manutenção da arquitetura e infraestrutura de dados, portanto, é responsável pela geração, projeto, construção e manutenção do ambiente de dados, sistema de processamento e armazenamento. A possível ausência da engenharia de dados afetaria a ciência e análise dos dados.

As principais responsabilidades da engenharia de dados estão relacionadas ao projeto e desenvolvimento de rotinas e objetos usados para armazenar dados, normalmente a rotina é o processo de carregamento (ETL) ou ingestão de dados. Incluso nesses procedimentos também incluem procedimentos para processamento, limpeza e qualidade de dados.

A ingestão de dados é um termo que significa que os dados são inseridos em estruturas de dados normalmente não relacionais, essas estruturas podem ser baseadas em sistemas de arquivo de código aberto.

Outra função principal e praticamente exclusiva de um engenheiro de dados é a criação, manutenção e operação de pipelines de dados. Os data pipelines são métodos de orquestração da movimentação de dados entre sistemas diferentes, cujo formato de armazenamento e gerenciamento do dado geralmente difere da fonte para o destino.

O pipeline de dados permite então a integração entre diferentes bases de armazenamento dos dados, pela automatização do processo de cópia de um local para outro com a aplicação de

transformações e consolidação dos dados movimentados, que podem ser disponibilizados em repositórios de armazenamento com alta performance e disponibilidade, como um data lake(lago de dados).

Os pipelines são gerenciados pelo engenheiro de dados, que atua em todo o ciclo de vida do processo, que consiste na definição de origem e destino e suas ligações, o fluxo dos dados caracterizado pelo ETL, os modos de armazenamento para os resultados das fases do pipeline, o modo de processamento utilizado, o fluxo de trabalho com as tarefas agendadas e automáticas, a monitoria de execução dos procedimentos e toda a stack de tecnologias empregadas em cada ciclo.

2.4.1 Tipos de processamento de dados em pipelines

Os pipelines para transpor de dados entre fontes diversas possuem paradigmas de processamento e execução a depender do seu propósito. Alguns sistemas de big data analytics necessitam de dados em tempo real, outros precisam de alimentação de dados esporadicamente, e tem alguns que necessitam de movimentação de dados em pulsos de duração quase instantânea. As formas de processamento do pipeline de dados são divididas nos seguintes modelos:

Batch Processing(Processamento em lotes): este modelo consiste no processamento de grandes volumes de dados agrupados de uma só vez, e várias tarefas de processamento em batch podem ser executada simultaneamente sem interrupções e em ordem sequencial. São apropriados para movimentação transacional de dados que não dependem de continuidade, como o processamento de pagamento de faturas ou a carga de sistemas analíticos.

Real-time processing: processamento de dados em tempo real é característico de sistemas reativos. O tempo de resposta na interação entre uma origem e um destino de dados deve ser um pulso de intervalo que tenda a zero na medição de tempo real. Processamento de tempo real é designado para trocas de dados em que a atualização entre a entrada e saída precisa ser simultânea, como uma transação bancária em um caixa eletrônico ou um controlador de voo.

Streaming processing: o processamento corrente ou stream processing trata-se de uma troca de dados praticamente instantânea entre origem e destino, em que o intervalo de tempo de duração da interatividade e do transporte dos dados precisa ser relativamente curto e o tempo de processamento dos dados de entrada na fonte precisam ser iguais aos de saída no destino. O

processamento em stream é aplicado a trocas de dados entre sistemas que precisam de uma sincronia com baixo nível de atraso, como serviços de vídeos sob demanda ou portais analíticos on-line.

2.5 BUSINESS INTELLIGENCE

Business intelligence é “Um conjunto de conceitos, métodos e recursos tecnológicos que habilitam a obtenção e distribuição de informações geradas a partir de dados operacionais, históricos e externos, visando proporcionar subsídios para a tomada de decisões gerenciais e estratégicas”, segundo o Gartner Group. Traduzido como “Inteligência de negócios”, é um conjunto de ferramentas e metodologias de aplicação de softwares diversificados na exploração e análise de dados de um negócio com objetivo de apoiar tomadas de decisão empresariais, sendo uma forma automatizada e de instrumentalizar o processo decisório do negócio através das TIC. Baseia-se na capacidade e potencial analítico que integram em um único lugar toda e qualquer informação necessária nos processos de decisão.

2.5.1 Arquitetura de BI

A arquitetura de Business Intelligence é uma estrutura organizacional de gerenciamento de informações e componentes integrados que são utilizados para a construção dos sistemas de Inteligência de negócios e as tecnologias empregadas nestes.

Os componentes compreendem as estruturas de ferramentas aplicadas, como as fontes de dados utilizados por analistas usuários do sistema de BI para avaliar aspectos da empresa que apoiem a tomada de decisões assertivas e a transformação de dados brutos obtidos destas fontes e das informações contidas neles em conhecimento referenciado para o uso dos processos de BI.

As arquiteturas dependem do escopo e do propósito de aplicação do BI, no entanto seguem determinados padrões, que compreendem na organização dos conceitos chave envolvidos como fonte, transformação, armazenamento, processamento e análise dos dados. A maioria das arquiteturas utilizam o esquema presente na imagem a seguir:

Nesta imagem, nota-se o relacionamento entre as fontes de dados distintas que são entregues no processo de extração, transformação e carregamento (ETL) dos dados para organizar as informações neles contidas, que é a etapa de preparação e adaptação dos dados

para então armazená-los em um banco de dados para big data, construído em uma modelagem de Data Warehouse ou Data Mart, a depender do propósito e escopo. Por fim, os dados contidos em um repositório para armazenagem massiva são utilizados na aplicação dos processos finais de Business intelligence, que compreendem a análise de dados caracterizada pelo processamento analítico de dados on-line (OLAP) e a ciência de dados que se caracteriza pelas técnicas de mineração de dados (Data Mining).

Uma típica arquitetura de soluções de BI segue uma série de três etapas, conhecidas como camadas estruturais. São estas:

ETL(Extract Transform Load): conjunto de aplicativos e ferramentas que fazem a coleta de dados nos repositórios da organização, procedem com a limpeza e transformação para enfim carregar o DW (BRACKET, 1996; INMON, 1997; KIMBALL et al., 1998);

Repositório de dados (Data Warehouse ou Data Marts): repositório de dados integrado e não-volátil onde são armazenados os dados transformados pelo módulo ETL. Esse repositório deverá suportar as demandas analíticas das ferramentas de apoio à decisão e os aplicativos de extração de conhecimento (INMON, 1997; KIMBALL et al., 1998; SELL; PACHECO, 2001);

Apresentação de dados (Front-end): diz respeito ao conjunto de instrumentos que serão utilizados pelos usuários na organização para navegar no DW. Esses instrumentos correspondem a relatórios previamente configurados, aplicativos para confecção de relatórios, ferramentas OLAP (On-line Analytical Processing), ferramentas de Data Mining (mineração de dados), entre outras (BERRY; LINOFF, 1997; BERSON, 1997; GONZAGA; 2005; THOMSEN, 2002).

2.5.2 Data-Driven-Design (DDD)

A abordagem Data Driven Design consiste em desenvolver e projetar soluções baseada em dados, ou seja, arquitetar ações e estabelecer resposta a questões de negócios orientadas aos dados disponíveis e analisados. Dados são características e atributos atômicos a respeito de itens do mundo real, que se referenciados geram informação, quando ganham contexto. Esta contextualização pode ser aplicada e embasar a tomada de decisões para tomar ações resolutivas, de planejamento e arquitetura de processos de negócio nos mais variados segmentos. Esta base para tomar decisões de negócio ancorada nos dados a respeito deste é o que caracteriza a prática chamada de Projeto Orientado ao Dado ou Data-Driven-Design.

2.6 ANÁLISE DE DADOS

A análise de dados é o trabalho de compreender uma grande quantidade de dados não estruturados que precisam ser compilados e organizados. Nesse processo, os profissionais podem obter resultados sobre vendas, marketing, relacionamento com clientes e outras possibilidades. (FERREIRA, 2019)

A análise de dados é um esforço para aprofundar os dados coletados pela empresa em suas principais fontes. E-mails, plataformas de gerenciamento, ferramentas automatizadas, planilhas, documentos e muitas outras fontes geram grandes quantidades de material não estruturado, mas muito útil, conhecido como big data.

Após a compilação, uma grande quantidade de informações pode ser analisada tecnicamente com o auxílio da tecnologia.

Existem sistemas de suporte aos analistas que podem identificar automaticamente os padrões de comportamento desses dados.

A análise dos dados pode garantir que, partindo da matéria-prima, trabalhem sempre com a tecnologia e entendamos a empresa sob diversos ângulos com o suporte de tecnologia de ponta.

2.6.1 Tipos de análise de dados em Big Data

A análise de dados em Big data é definida mediante a sua finalidade, estrutura organizacional dos dados e foco analítico. Os fins da análise podem ser previsão, prescrição, diagnóstico, a estrutura pode ser composta por dados históricos, conjuntos de associações, variáveis numéricas contínuas e as análises podem focar em dados do passado, tendências ou correlação entre dados. Cada tipo de abordagem possui as características supracitadas que os compõe, são eles:

Análise preditiva: possui a finalidade de predição de comportamentos futuros, com foco em situações específicas que condicionem determinados acontecimentos através de padrões.

Análise prescritiva: destina-se à avaliação consequencial de ações tomadas com a análise dos dados provenientes de uma ação. Desta forma pode-se traçar prescrições de rumos e medidas necessárias para atingir resultados específicos.

Análise descritiva: visa analisar de maneira geral os fatos que ocorrem no mundo real em tempo presente, sem aferir juízo de valor analítico, restrita apenas a dispor a visualização e exibição dos dados em ocorrência.

Análise diagnóstica: pauta-se na análise de impacto de ações tomadas ou da ocorrência de eventos, a partir do levantamento de dados correlatos do passado, para que seja possível traçar causalidade entre eles.

2.6.2 Análise de dados Ad Hoc

A expressão Ad hoc é latina e significa "para este fim". Ou seja, a consulta é criada para um propósito específico apenas para atender a necessidades específicas em um momento específico.

Ad Hoc é uma espécie de consulta SQL no banco de dados, que é criada dinamicamente de acordo com necessidades específicas (não pode ser generalizada, ou seja, pode ser utilizada em mais de uma situação) quando há demanda, ao invés de ser criada instantaneamente e salvos no SGBD para que possam ser reutilizados posteriormente.

Esse tipo de consulta é causado por solicitações não planejadas e imprevistas.

Normalmente, as solicitações não planejadas são características de aplicações de suporte à decisão, como aplicações de BI ou Data Mining, portanto, as consultas ad hoc são muito comuns em tais sistemas. As consultas temporárias são projetadas para fins específicos e são diferentes das consultas predefinidas.

Consultas temporárias podem ser criadas de várias maneiras, por exemplo, diretamente no processador de linguagem de consulta / script, criado manualmente, linha de comando, gerado dinamicamente na aplicação, entre outros.

Consultas temporárias não são armazenadas no banco de dados e o mecanismo do banco deve sempre analisar, criar planos de execução, entre outros. Quando a consulta é executada, eventualmente, elas se tornam consultas não otimizadas.

Se foi utilizado uma string dinâmica para criar uma consulta temporária, será uma consulta dinâmica.

As consultas dinâmicas são consideradas consultas ad hoc porque também não contêm instruções SQL criadas como procedimentos armazenados.

O problema que pode surgir ao usar consultas ad hoc é que elas podem ser suscetíveis a ataques de injeção SQL, pois quando esse tipo de consulta é executado, às vezes os objetos do

banco de dados são eventualmente expostos. Portanto, neste caso, será melhor utilizar parâmetros ao criar as consultas, pois isso ajuda a proteger os objetos do banco.

2.7 Ciência de dados

O termo ciência de dados é novo, pois surgiu por volta de 1960, podemos dizer que a ciência de dados é um tipo de ciência nova, por esse motivo ela pode ser mal compreendida. A ciência de dados procura examinar os dados em todo seu curso de vida. (AMARAL, 2016)

A importância da Ciência de dados está atrelada a ela permitir a extração de informações bastante valiosas a partir de uma base de dados.

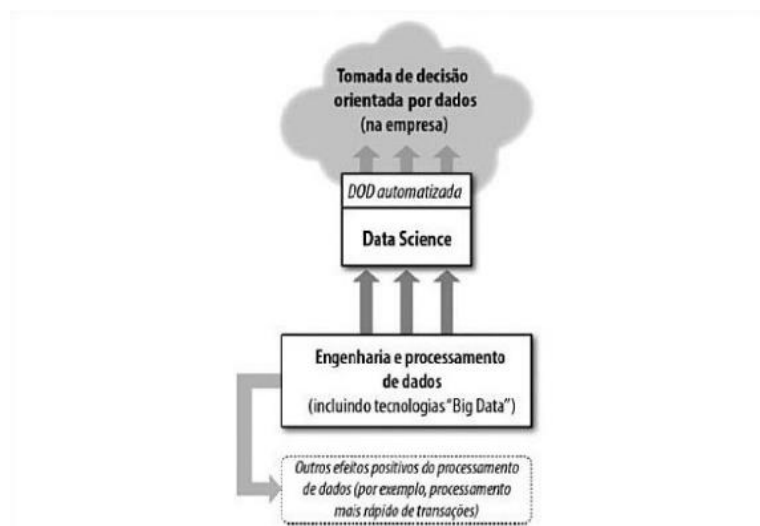
A ciência de dados é um campo muito auspicioso para ser estudado e analisado, e a exploração desse gere grandes volumes de dados advindo de diversas fontes.

Há várias disciplinas envolvidas na ciência de dados. Sendo elas:

- Estatística
- Computação
- Conhecimento do negócio
- Matemática

O processo se inicia com a coleta dos dados mediante a um questionamento correto sobre o problema e objetivo, após esse passo é feito a análise dos dados, utilizando a visualização, aplicação de técnicas e algoritmos, termina com a comunicação dos resultados. (PASSOS, 2016)

Figura 13 - Ciência de dados no contexto dos diversos processos relacionados a dados na organização.



A Figura 13 mostra a ciência de dados na circunstância de inúmeros outros processos correlacionados e associados com dados na organização. A imagem diferencia ciência de dados de outras perspectivas do processamento de dados.

2.8 MINERAÇÃO DE DADOS

A mineração de dados (em inglês, Data mining) é um processo analítico de extração de conhecimento através da exploração automatizada de uma base de dados, com objetivo de detectar anomalias, padrões, correlações e potenciais causalidades entre os dados transacionais armazenados em um banco de dados padrão ou até mesmo em um conjunto de grande volume de dados (em inglês, Big Data). O objetivo principal que pode ser alcançado a partir disto é a descoberta de conhecimento ou sabedoria em uma base de dados, após sua estruturação para ser minerada.

As técnicas de mineração são algoritmos que analisam uma base ou conjunto de dados para aplicar sua metodologia para busca de conhecimento e produção de sabedoria de forma automática e preditiva, ao contrário das técnicas convencionais de *business intelligence*, que consiste em definir hipóteses, problemas e estabelecimento de indicativos para serem testados e avaliados contra a base de dados através da análise de dados passados, apresentando os resultados obtidos na coleta para auxiliar na tomada de decisão orientada a dados.

Desta maneira, pode-se aferir que a mineração de dados é uma abordagem proativa na descoberta de conhecimento em base de dados com foco em projeção futura e generalista de um domínio de negócios a partir dos dados coletados, além da simples identificação de tendências anteriores contidas na base. Sendo um braço técnico da ciência de dados, trata-se da aplicação de inteligência computacional com enfoque em análise exploratória dos dados particulares de um segmento de mercado para produzir conhecimentos específicos sobre este setor.

2.8.1 A pirâmide do conhecimento para mineração de dados: Hierarquia DIWK

Figura 14 - A pirâmide de conhecimento



Fonte: towardsdatascience.com

A pirâmide de conhecimento conhecida como Data-Information-Knowledge-Wisdom(do inglês, dado-informação-conhecimento-sabedoria) é uma hierarquia de conceitos muito utilizado em análise e ciência de dados, e também na arquivologia e na ciência da informação.

Foi primeiramente pesquisado pelo teórico Russel Ackoff por volta da década de 1980. Russel aponta que um dado em isolado não possui significado em si, pois necessita de interpretação para ganhar contextualização e referência sendo transformado em informação e o conjunto de compreensões a cerca da informação é entendido como conhecimento. A sabedoria, no topo da pirâmide, seria o resultado composto pelas visões obtidas através dos diferentes cenários de conhecimento analisados.

Para Jennifer Rowley(2007), as informações são tipicamente definidas em termos de dados, o conhecimento em termos de informação e a sabedoria em termos de conhecimento.

Como cada etapa é uma composição da seguinte, é necessário analisar cada esfera da pirâmide para agrupar o nível e construir o caminho para o seguinte. Desta forma, a análise de dados é iterativa e ultrapassa etapas até chegar no momento desejado. Certas análises preocupam-se com produção de informações, outras de conhecimento e algumas buscam uma sabedoria acerca dos cenários. Em todos os casos, o dado é o objeto de estudo e a finalidade é sempre alguma etapa subsequente. Cada um dos estágios da pirâmide pode ser definido da seguinte maneira:

- Dado: dados são registros factuais de situações do mundo real, que produzem sinais ou símbolos, quantitativos ou qualitativos e nada a mais que isto. Podem ser

compreendidos como ruídos, eventos, registros específicos ou qualquer ocorrência sem significado intrínseco. São inertes, desorganizados e não processados.

- **Informação:** surge quando um ou mais dados são contextualizados e ganham referências, tornando-os úteis para um propósito. Ao aplicar classificação, organização e estruturar o(s) dado(s) produz-se informação, que é orientada a responder perguntas com os dados.
- **Conhecimento:** é o resultado da estruturação e organização de toda informação produzida a partir de dados processados, para atingir objetivos e aferir utilidade as informações estruturadas. Está relacionado implicitamente a aprendizado e permite a detecção de padrões informacionais e estabelecer modelos preditivos.
- **Sabedoria:** é a interpretação geral de todo conhecimento produzido, que ultrapassa a fronteira da detecção de padrões e escala para a explicação e entendimento do que os motiva. É a forma de compreender o conhecimento sobre informações do passado para projetar com assertividade tendências futuras.

Neste trabalho que se segue, o objetivo final através da hierarquia DIKW é a produção de conhecimento acerca da base de dados do Enem 2019 através da análise de dados ad-hoc para extrair informações estatísticas a cerca dela e a extração de sabedoria a partir do conhecimento contido nesta base, produzindo modelos probabilísticos para detecção de correlação e confiança na inferência de causalidade entre eles, através da mineração de dados pela descoberta das regras de associação sob a ótica operacional do ciclo KDD.

2.8.2 KDD

O processo KDD (do inglês, Knowledge Discovery in Databases, ou Descoberta de Conhecimento em Bancos de Dados) objetiva na alise de informação contextualizada de uma base ou banco de dados que contenham um alto volume de armazenamento (Big Data, do inglês, Mega-Dados).

O objetivo deste padrão analítico processual é orientar o fluxo de tratamento de dados em larga escala, organizando uma sequência cíclica da mineração de dados que leve em consideração todas as partes deste processo, abrangendo as várias áreas ou corpos de conhecimento por trás da manipulação de dados como infraestrutura de armazenamento (bancos de dados, data warehouses, data lakes), ciência de dados (estatística, matemática aplicada.

Inteligência artificial) e análise de dados (detecção de padrões, testes de hipóteses, visualização de dados).

O KDD é uma estrutura em ciclo comum a qualquer técnica de mineração de dados e vai envolver todas estas áreas subordinadas a esta abordagem analítica orientada a dados (do inglês, data-driven analysis). As etapas do processo de descoberta de conhecimento em bancos de dados são:

Coleta (ou seleção) de dados: uma vez estabelecido o entendimento do negócio e delimitado o domínio de risco pertencente a ele onde se deseja produzir o conhecimento, é necessário coletar os dados de uma ou várias fontes, através da seleção daquelas de maior relevância para o objetivo da análise e armazenamento dos conjuntos de dados obtidos das bases ou repositórios escolhidos. O objetivo a alcançar é selecionar conjuntos de dados úteis e então disponibilizá-los para o pré-processamento.

Pré-processamento e limpeza dos dados: nesta etapa, realiza-se o carregamento inicial para uma análise preliminar da situação dos dados. Este procedimento visa varrer todo o conjunto ou base de dados para identificar possíveis inconsistências, incompletudes, heterogeneidade na integração de dados distintos, incompletude e redundância sem propósito que incida em cada dado. Também se busca potenciais problemas que o formato ou disposição de um dado possa ocasionar durante a aplicação do algoritmo no momento da mineração. O Objetivo é ajustar inconsistências, mitigar falhas de integridade e descartar dados inúteis para a técnica a se aplicar.

Transformação de dados: feita a remoção de dados desnecessários ou ofensores no pré-processamento dos dados coletados, realiza-se a transformação dos que foram aproveitados para que sejam compatíveis com o algoritmo de mineração a ser usado. Uma vez que cada algoritmo trata dados de formas específicas, é preciso compatibilizar o formato do dado para que seja adaptado a forma como a mineração será realizada. Leva-se em conta o tipo de dado que é suportado, se é qualitativo ou quantitativo, e o formato do dado, se é discreto ou contínuo.

Mineração de dados: é o objetivo principal e a etapa mais importante do ciclo KDD, pois é nela que se aplica ferramentas automatizadas e algoritmos para descoberta de correlações, causalidades, padrões e tendências entre informações de uma base de dados, a fim de prever um conhecimento ou sabedoria a partir dos resultados desta análise de dados.

Avaliação e interpretação dos resultados: etapa final de um ciclo KDD, onde interpreta-se todos os resultados obtidos da mineração e, a partir disto, conclui-se a produção de conhecimento extraída desta verificação dos dados minerados.

2.8.3 Técnicas de mineração de dados

Cada base de dados possui um formato apropriado a um tipo de técnica de mineração a ser escolhida, que também tem seu propósito mais adequado em termos de resultado obtido. Algumas técnicas visam avaliar bases de dados transacionais, ou seja, tratar cada linha ou registro como uma ocorrência de fatores. Pode-se também analisar uma base de dados que tenha atributos qualitativos ou quantitativos e, ao levar em conta este último, ainda pode-se distinguir a análise de valores discretos ou contínuos. Estas formas de implementação do DataMining têm objetivos específicos, e para cada qual determinadas ferramentas ou algoritmos especializados. As mais relevantes são:

Descoberta de associações ou análise de regras de associação: Consiste na varredura de uma base de dados que visa buscar relações entre eles para prever padrões de correlação e aferir causalidade entre cada dado, através da avaliação probabilística de regularidade associativa entre cada conjunto de valores analisado. O funcionamento do algoritmo é dado pela probabilidade da união de conjuntos e pela frequência de ocorrência de cada item em uma base transacional de dados, comparando a quantidade de cada valor em um conjunto de dados e a probabilidade de ocorrer em concomitância com outro de forma dependente.

Classificação: Busca obter, através de modelos preditivos, conjuntos de modelos descritivos discretizados de uma base de dados de acordo com conceitos pré-definidos conhecidos como classe, com objetivo de utilizar essas classificações para prever novas entradas de dados ainda não rotulados, de modo a classificá-las de acordo com as classes de objetos estabelecidas. Como esta técnica opera através de modelos pré-estabelecidos de rótulos, seu algoritmo utiliza-se de aprendizado supervisionado de máquina.

Agrupamento ou Clusterization: Técnica similar às análises preditivas, mas com sentido inverso, pois emprega a análise de dados brutos sem predefinição de rótulos para então identificar agrupamentos (do inglês, clusters) dos objetos discretos que definem uma classe. Desta forma, um algoritmo de clusterização aglutina dados com similaridades para criar grupos a serem rotulados posteriormente. Devido ao fato de aplicar análise sem orientação preliminar, o algoritmo de agrupamento emprega aprendizado não-supervisionado.

Regressão: Trata-se de uma análise preditiva que visa analisar um conjunto de valores como dado entrante e prever valores de saída a partir deles. É conhecida como predição funcional, pois associa atributos discretos a atributos contínuos em uma relação de dependência

funcional e por ter como objetivo a previsão de atributos contínuo, isso difere os modelos de regressão dos de classificação. Como estes algoritmos analisam entradas dependentes e resultam em regras funcionais preditivas, utilizam aprendizado não supervisionado.

Análise sequencial: É uma abordagem cronológica orientado a eventos, que analisa o intervalo temporal da ocorrência de determinados itens em uma base de dados dada uma ordem pré-definida. Esta técnica tenta prever padrões sequenciais de valores em uma base de dados. Por realizar uma busca em um conjunto de dados baseada em um padrão pré-definido, o algoritmo sequencial aplica aprendizado supervisionado.

Deteção de Ruídos ou Outliers: Esta técnica tem como objetivo avaliar dados a partir de padrões pré-definidos para detectar transações que fujam deste padrão, sendo anomalias na base da dados (ou outliers). Estes desvios são conhecidos como ruídos, por destoarem dos padrões e regularidades em um conjunto de dados, sendo frequentemente valores descartados em outras técnicas, no entanto a abordagem de detecção foca justamente nestas discrepâncias em bases de dados. Por partir de padrões já conhecidos para analisar a base, o algoritmo de detecção de anomalias(ruídos) emprega aprendizado supervisionado.

2.8.3.1 Técnica de mineração e algoritmo escolhidos

Foi escolhido para o propósito deste trabalho a técnica de mineração de dados por descoberta de regras de associação, que utiliza uma abordagem de probabilística para avaliar correlação entre conjuntos de itens em uma base de dados transacional para aferir a probabilidade de ocorrerem em conjunto e de um conjunto implicar na ocorrência simultânea de outro, sendo um vetor analítico de causalidade.

O algoritmo de regras de associação aplicado foi o Apriori Association rules, disponibilizado como biblioteca na linguagem R, que realiza uma varredura recursiva de um conjunto de dados para mapear transações e partir delas, descobrir as regras de associações e identificar suas medidas.

2.8.3.1.1 Definição formal do Algoritmo de Mineração por Regras de Associação (ARM)

Seja $I = \{I_1, I_2, \dots, I_m\}$ um conjunto de m atributos(dados) distintos. Seja D uma base de dados, onde cada registro ou tupla T possui um identificador único e contém um conjunto de itens tais quais $T \subseteq I$. Uma regra de associação é uma implicação na forma de $X \Rightarrow Y$, onde X ,

$Y \subset I$ são conjuntos de itens chamados de itemsets e $X \cap Y = \emptyset$. O atributo X é definido como o antecedente e o atributo Y é definido como consequente.

O **suporte(support)** de uma regra de associação é a razão (em porcentagem) entre os registros que contém $X \cup Y$ e o número total de registros na base. Para um dado número de registros, a **confiança (confidence)** é a razão (em porcentagem) entre número de registros que contém $X \cup Y$ e o número de registros que contém somente X . Seja a confiança **C** da regra $X \Rightarrow Y$ e o suporte **S** de Y , o parâmetro **alavancagem (Lift)** é uma razão entre $C(X \Rightarrow Y)$ e $S(Y)$.

O parâmetro support mede a frequência em que conjuntos de itens de uma regra ocorrem em união. Já o confidence mede a probabilidade da união entre conjuntos X e Y onde $X \Rightarrow Y$ em relação a probabilidade de Y . O Lift mede a relação entre a confidence de $X \Rightarrow Y$ e o support de Y .

2.8.3.1.2 Técnica e algoritmo escolhido: regras de associação Apriori (*Apriori Association Rules*)

Os algoritmos de regras de associação atuam em bases de dados em formato transacional e com dados quantitativos discretos ou qualitativos, de modo que sejam contabilizados em termos de repetição ou frequência de aparição nas transações em um conjunto de dados analisado, a fim de estabelecer causalidade entre cada item de um conjunto.

Uma transação é um registro em uma base de dados que contém uma identificação e é composta por um conjunto de itens armazenados que permitem caracterizá-la pela combinação destes elementos, sendo definida pela relação entre identificador e conjunto de itens. A técnica de descobertas de regras de associação pode ser aplicada apenas em bancos ou bases de dados em formato de transação, pois avalia probabilidades de aparição simultânea de itens em um dado conjunto, por isto avalia cada registro transacional quantificando seus atributos ou itens.

A técnica “apriori” atua nestes conjuntos varrendo iterativamente cada transação em busca da repetição de cada item para quantificar sua frequência em relação ao todo e testar associação entre os itens presentes, o que resultará em uma regra de associação entre os dados. Ao aplicar uma abordagem iterativa, o algoritmo “apriori” analisa a frequência de conjuntos de itens de tamanho k baseado na frequência dos itens de tamanho $k-1$, que se baseia no cálculo da frequência dos conjuntos de tamanho $k-2$ e assim sucessivamente até avaliar cada transação registrada.

A frequência simultânea de todos os itens presentes em uma determinada regra é chamada de suporte (support), que se dá pela probabilidade de ocorrerem em conjunto em toda a base, e as relações causais na relação entre cada item dela chama-se confiança (confidence). A confiança é obtida através da probabilidade da ocorrência simultânea de itens e de um conjunto deles ocasionar um outro, o que estabelece o percentual de dependência entre um item e outro. A hipótese associativa que mede a relação entre a frequência de um item consequente e a probabilidade de consequência denomina-se elevação (lift). O lift é a razão entre confiança e suporte de uma regra, que determina se a correlação entre os itens avaliados é positiva, negativa ou nula.

As regras de associação são pautadas em análise de correlação e causalidade. As técnicas da descoberta de associações avaliam ocorrência de itens concomitantes em bases transacionais, a probabilidade da ocorrência de certos itens em condição de outros e, existindo o condicionamento, se o sentido é positivo ou negativo. Estes parâmetros são os que determinam a validade ou relevância de uma regra de associação específica, sendo representados pelo suporte, confiança e elevação, respectivamente. Em qualquer algoritmo de associação, as regras serão geradas a partir dos valores mínimos destes indicadores estatísticos pré-definidos, que são medidores probabilísticos determinantes para descarte ou aceitação de cada regra e para o teste de hipótese que as embasa.

Cada conjunto de dados dentre um registro transacional é tido como um “conjunto de dados candidato”, pois pode vir a se tornar uma regra associativa caso os valores de confiança e suporte atendam ao mínimo estabelecido arbitrariamente, cujo a hipótese de correlação é validada pelo lift ou alavancagem, que indicará o sentido da regra. A forma como as regras candidatas são detectadas e descobertas depende do algoritmo escolhido.

Em se tratando do “Apriori”, o esquema utilizado para varrer toda a base para descoberta de regras consiste em analisar cada conjunto de dados a partir de tamanhos unitários e, a partir deles, avaliar conjuntos maiores de forma incremental e iterativa, pois junta cada conjunto prévio para transformá-lo em um maior e elimina os menores que tiverem frequência abaixo do valor definido. Desta forma, o algoritmo apriori reduz os conjuntos de dados considerados para exploração apenas aqueles que têm o suporte maior do que o valor mínimo encontrado em todo o conjunto de dados.

Escolheu-se a técnica de descoberta de associações e a utilização do algoritmo de regras de associação apriori, pois possibilita listar todos os conjuntos de itens de diferentes tamanhos

no banco de dados que possuam frequência suficientemente consideráveis e, ao adotar uma abordagem iterativa, possibilita uma varredura eficiente de uma grande base de dados.

Como o objetivo deste trabalho é correlacionar fatores socioeconômicos e característicos aos participantes do ENEM ao seu desempenho e participação neste e traçar perfis padrões entre todos eles de acordo com estes critérios, descobrir associação entre estes itens torna-se a forma mais adequada de aferir causalidade a itens relacionados, o que também viabiliza descobrir relações entre os dados cadastrais em si que sejam aproveitáveis no desenho de um perfil social, econômico, geográfico ou até mesmo psicopedagógico de um candidato, porém este último foge do escopo atual deste trabalho.

A mineração de dados por regras de associação demonstra-se uma ferramenta poderosa de apoio na análise social do ENEM e também pode ser útil no estudo da educação inclusive no exame, uma oportunidade para trabalhos futuros. Neste trabalho que se segue, submete-se a base de dados ao ciclo do KDD e busca-se entender o exame nacional do ensino médio pela perspectiva pedagógica para analisar aspectos sociais, psicológicos, econômicos e geográficos a respeito dos participantes e como estes fatores podem afetar o seu desempenho nele, avaliar o impacto destes critérios no aproveitamento do aluno em cada área de conhecimento, nas competências da redação e na sua participação nas provas, ausência e possíveis eliminações.

A partir desta avaliação estatística computacional dos microdados do ENEM, pode-se revelar conhecimento acerca do perfil social de quem realiza a prova extrapolar a interpretação dos resultados para descobrir sabedoria a respeito da situação do ensino médio regular no país, pois o exame é estruturado pautado em uma base comum curricular a toda cadeia de ensino secundário no Brasil.

2.9 ETL & ELT

Sigla do inglês “Extract Transform Load”(Extrair, transformar e carregar), é o processo de ingestão e carga de dados de uma ou várias bases de dados através da integração diferentes sistemas para compor um conjunto de dados, que será consumido nos processos de *business intelligence* e *data science*. Este procedimento é realizado por softwares especializados em integrar diferentes fontes de dados para se extrair o tipo de dado que provém informações úteis ao modelo a ser construído para exploração ou análise. As características de cada etapa do ETL (no português, ETC) são:

Extract(Extração): é a fase de integração multi-sistemas para capturar os conjuntos de dados que serão utilizados no destino para a análise

Transform(Transformação): etapa que engloba a preparação, limpeza, pré processamento e adequação dos dados para o uso no destino, de forma que todo dado seja compatível com as técnicas de BI utilizadas, tanto em tipos como em formatos.

Load(Carregamento): o processo final, após as duas fases iniciais de preparação do conjunto de dados, onde é feito o carregamento do modelo de dados resultante da transformação para os destinos de armazenamento, podendo ser data warehouses, data marts, data lakes e afins.

2.9.1 ELT

O processo ELT vem da sigla em inglês para Extract Load and Transform e pode ser considerado uma modernização do ETL tradicional, que permite uma agilidade e maior eficiência no processo da ingestão das fontes de dados e armazenamentos no destino ao inverter a abordagem do fluxo processual no tratamento dos dados.

No ELT, é realizado primeiro a extração dos dados brutos de uma ou várias fontes, estes dados são carregados sem quaisquer alteração ou modificação para um destino de armazenamento, para aí então sofrer transformações. O objetivo com esta inversão de etapas é diminuir substancialmente o tempo de carga entre base de dados da fonte e o conjunto de dados do destino.

Esta inversão permite com que os dados sejam disponíveis em formatos mais rápidos para o transporte, o que torna a velocidade de transferência no carregamento mais acelerada e coloca a transformação na ponta destinada ao modelo do dado modificado. Para situações em que é necessário aplicar transformações muito específicas, em grande quantidade e que gere dados em formatos mais robustos, o ELT é uma opção mais adequada.

2.10 MODELO DE PROCESSAMENTO DE DADOS EM BANCOS DE DADOS

Os bancos de dados possuem dois formatos operacionais e de aplicação, que são os transacionais e analíticos. Bancos de dados transacionais caracterizam-se pelo funcionamento baseado em interações entre usuários e banco caracterizada pela manipulação de dados concorrente e mutuamente exclusiva entre usuários, interação esta referenciada por um registro

com histórico rastreável, fechamento de recursos concorrentes(lock) e em alguns casos a possibilidade de desfazer a operação(rollback/undo).

Já os bancos de dados analíticos, são concebidos para armazenar e prover métricas sob medida conforme a demanda de analistas de dados que consultem o banco, e tais medidas são organizadas em modelos geralmente multidimensionais, ou por vezes relacionais(tabulares). A principal característica dos bancos analíticos é a interatividade entre o usuário e banco de forma concorrente e consultiva, com retorno da métrica solicitada e não a manipulação dos dados.

Os modelos de processamento dos dados em um banco são divididos com base nestes dois modelos, sendo caracterizados como On-line transactional Processing (Processamento transacional online) e On-line Analytical Processing (Processamento analítico online).

2.10.1 OLTP

O Processamento Transacional On-line refere-se a sistemas de bancos de dados transacionais, que são aqueles que fazem parte da área operacional de um negócio. Os bancos de dados orientados a transação provêm armazenamento de bases de dados de uma empresa para que possam ser manipulados e registrem informações críticas para os negócios de uma corporação. Focam em realização de um vasto número de transações em tempo real em uma base de dados conectados em aplicações de linha de frente de um negócio, como CRMs, ERPs e E-Commerce.

2.10.2 OLAP

O Processamento Analítico On-line está relacionado a sistemas de bancos de dados analíticos, que são os que fazem parte da área estratégica de um negócio. Os bancos de dados orientados a análise destinam-se a coletar os registros de dados e informações contidos nos bancos transacionais e aplicar a extração de métricas e testes de hipóteses com base nestes dados, a fim de produzir conhecimento sobre a base através da análise realizada, com intuito de servir de apoio estratégico na tomada de decisão para os negócios de uma organização. Focam em aplicar análises de alta velocidade em grandes volumes de dados provenientes de data warehouses, data marts e data lakes.

2.11 MODELO DE ARMAZENAMENTO DE DADOS MASSIVOS (BIG DATA)

Sistemas de informação funcionais apresentam dificuldades em integrar informações, especialmente no que tange a necessidade de análise de histórico de problemas ou situações ou então analisar tendências para elaborar modelos preditivos. Além disto, também há problemas de redundância nos dados e, conforme a escala do sistema quando possui patamar de dados massivos (Big Data) elevam a acentuação da gravidade destes problemas.

2.11.1 Data Warehouses

Os armazéns de dados (do inglês, datawarehouses) ou DW são conjuntos de dados (ou datasets) específicos para um propósito que são extraídos de diversas fontes distintas como bancos de dados relacionais OLTP, CRMs, ERPs, Websites entre outros.

Os DWs são um tipo de base de dados centralizadora e coesa de informações para convergência de dados e informações com utilidade analítica para uma empresa, portanto é construído com a definição de um modelo empresarial que abrange a apresentação de suas principais entidades e relacionamentos críticos.

Segundo Bill Inmon, em seu livro “Building the Datawarehouse”, o Datawarehouse é um conjunto de dados orientado, orientado para o assunto, integrado, não volátil, variante ao tempo, no apoio de decisões gerenciais. Seu foco é a organização de todo o ecossistema de negócios como um todo, englobando várias áreas de assuntos distintos e consequentemente criando data marts para área em específico.

2.11.2 Data Mart

O mercado de dados (do inglês, data mart) é uma versão reduzida de um DW, que abrange somente uma determinada área de assunto para oferecer informações mais detalhadas sobre o mercado ou departamento em questão.

Podem obtidos através de sistemas de bancos de dados transacionais tradicionais ou a partir de um DW central como fonte, que agrega dados de diversos outros sistemas para si.

Buscando dados de um sistema ou fonte de dados transacional diretamente, é possível construir um DM mais rapidamente, por focar apenas no assunto específico que ele atende e não precisar tratar todos os outros envolvidos na estrutura de um DW.

2.11.3 Data lake

O lago de dados (do inglês, data lake) é um repositório central de uma empresa para armazenamento e disponibilização de todos os dados de valor gerados pela empresa, como imagens, documentos textuais, logs de chat (conversa por texto ou voz), vídeos, etc. São repositórios para dado de qualquer tipo e qualquer formato, estruturado, semi ou não estruturado e diferentes formatos de arquivo suportados.

O objetivo do DL é disponibilizar dados de diferentes fontes para serem consumidos por vários destinos, para fins de tratamento e análise diversificados. São fontes de dados brutos para transformação e processo analítico e também destino dos conjuntos de dados processados, transformados e treinados nos modelos onde se aplica análise exploratória, ad-hoc, mineração ou até mesmo machine learning.

Os data lakes são soluções ideais para casos em que se tem muitos dados de várias fontes diferentes com inúmeras possibilidades de extração de conhecimento e informações sobre eles, ou também quando não há um plano estruturado e bem definido do que fazer com uma quantidade considerável de dados e existe a necessidade de organizá-los em um lugar.

O nome Data Lake foi cunhado por James Dixon, CTO (Chief Technical Officer ou Administrador técnico) da Pentaho, solução tecnológica de BigData. A ideia por trás da terminologia é expressar o conceito de um local em comum de onde todos os dados vem e para todos os resultados vão, com níveis de gerenciamento e segurança bem definidos para separar cada tipo de dados de acordo com seu objetivo. Os princípios que caracterizam um data lake são: DW tradicional e moderno.

2.11.4 Modern Datawarehouse(MDW)

Os armazéns de dados tradicionais (do inglês, traditional datawarehouses) ou TDWs são organizados para aglutinar os dados em um único lugar de forma estática, e, portanto, precisa ser manualmente atualizado para atender as variações e evolução de um negócio e suas facetas distintas. Em situações de ganho de escalabilidade em alto nível, o custo de infraestrutura e manutenção pode ser elevado e não atender a demanda total de uma megaempresa.

Surge então a necessidade de implementar uma arquitetura moderna, sob demanda, mais rápida e que suporte mais tipos e formatos de dados diferentes. Neste contexto, os MDWs são

implementações que utilizam a estrutura do TDW de forma aprimorada e adaptada a estas mudanças e novos paradigmas para tratar os dados massivos.

A modernização dos DWs visa o suporte a dados não estruturados em estruturas não relacionais, com a tecnologia NoSQL(Não apenas SQL, do inglês Not Only SQL), utilizam serviços de computação em nuvem para aplicar auto escalonamento, alta disponibilidade e contingência através da internet e também são arquitetados para serem integrados aos DLs para consultarem dados diferentes de vários formatos e estruturas que não necessariamente são compatíveis entre si.

Desta forma, o MDW ultrapassa o SQL, modelo de entidade-relacionamento, as infraestruturas centralizadas de servidores dedicados e o OLAP, sendo soluções robustas e híbridas para atender a um mercado como um todo com suporte a qualquer tipo de implementação de BigData.

2.12 MODELAGEM DE DATABASES E DATAWAREHOUSES

A arquitetura de bancos de dados é um procedimento necessário para adequar o sistema construído as necessidades e regras de um negócio, em correspondência com suas necessidades do mundo real. A modelagem de um banco de dados divide-se em modelagem conceitual, que organiza visualmente as entidades de uma base, seus relacionamentos e suas características e é construído em diagramas, a modelagem lógica, que detalha regras de negócio e especificidades de cada ente em uma base, como tipos de dados e chaves e é constituído por tabelas e a modelagem física, que é a definição prática em código do projeto do banco e da aplicação das regras de negócio, onde são desenvolvidos os códigos SQL para a criação do banco de dados de fato. No contexto de bancos de dados relacionais, os modelos de construção são o modelo entidade e relacionamento para bancos OLTP tradicionais e os modelos star schema e snowflake, que são específicos para data marts e data warehouses.

2.12.1 Modelo e Diagrama Entidade-Relacionamento (MER / DER)

O MER é um paradigma de modelagem de sistemas de bancos de dados utilizado para descrever os objetos envolvidos em um negócio(entidades) com suas características(atributos) definidas pelos seus dados e as relações entre estes objetos(relacionamento).

As entidades são as partes interessadas envolvidas em um domínio de negócio, sendo classificados como físicos ou lógicos, a depender da sua forma de existência no mundo real. Entidades físicas são concretas, como produto, casa, pessoa, funcionário e as lógicas são abstratas como venda, salário, categoria de produtos, corrida de táxi, ou qualquer interação entre entidades concretas que gere um objeto não tangível.

Toda entidade é nomeada na forma de um substantivo de forma clara e atômica para expressar sua função dentro de um domínio. Exemplos de nomes de entidades são CategoriaProdutos, Venda, Aluno, PagamentoSalário, etc. Entidades são divididas em classificações segundo o motivo da sua existência, sendo elas:

- Entidade forte: são aquelas cujo a existência por si só é suficiente e independem de outras entidades para existir, como uma entidade produto, pessoa ou carro, por exemplo.
- Entidade fraca: são as que dependem da existência de outra entidade para poder fazer sentido em existir individualmente, como uma entidade venda associada a produto ou endereço associado a pessoa, por exemplo.
- Entidade associativa: surge quando há a necessidade de relacionar diretamente uma entidade com um relacionamento. As entidades somente se relacionam entre si, e o relacionamento expressa a relação entre elas, não sendo possível relacionar uma entidade a um relacionamento, então ao torná-lo uma entidade é possível relacioná-lo a outras. Geralmente surgem em relações de cardinalidade muitos para muitos, onde o relacionamento em si passa a agregar atributos de duas entidades. Em um exemplo em que se tem a entidade venda e produto e é possível ter muitas ocorrências nas relações de detalhes da venda, uma entidade relacionamento desconto pode ser criada para se relacionar com os detalhes da venda, o produto e a venda.

Os relacionamentos são os vínculos entre entidades, que são identificados por meio da sua cardinalidade, que mede a quantidade de objetos envolvidos em cada lado das partes da relação. Classifica-se pela sua cardinalidade da seguinte maneira:

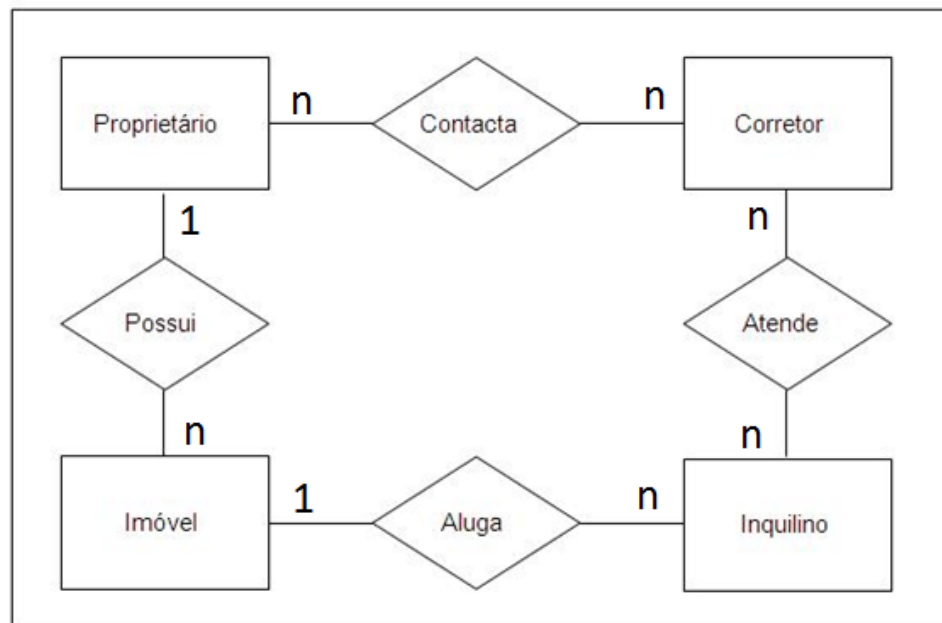
- Relacionamentos 1 para 1 (um para um): cada entidade envolvida referencia apenas mais uma entidade da outra e vice-versa
- Relacionamentos 1 para N (um para N): uma das entidades envolvidas pode referenciar várias unidades da outra, porém o outro lado pode referenciar apenas uma unidade.
- Relacionamentos 0 para 1 (zero para um): Uma das entidades envolvidas pode referenciar apenas uma unidade da outra, e esta outra parte pode referenciar uma ou nenhuma.

- Relacionamentos N para N (muitos para muitos): Uma das entidades envolvidas pode referenciar várias unidades da outra e vice-versa

Os atributos descrevem características que descrevem os dados de uma entidade. Podem ser descritivos, que expressam uma característica concreta, nominativos, que descrevem e identificam um objeto de forma única ou não, simples, quando são campos unificados e compostos, quando possuem vários campos.

Os diagramas de entidade e relacionamento (DER) concretizam a modelagem abstrato de dados que abrange as entidades e relacionamentos, e expressa suas interações, a cardinalidades e atributos envolvidos.

Figura 15 - Diagrama Entidade Relacionamento de sistema de imobiliária



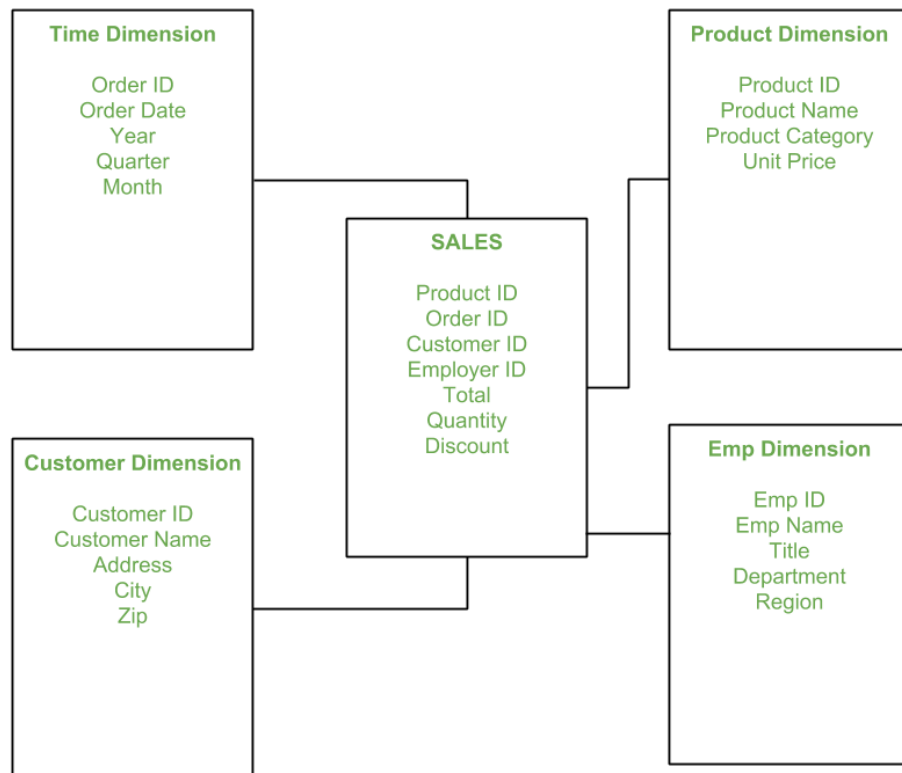
Fonte: devmedia.com.br

2.12.2 Modelagem de Star Schema e Snowflake

O esquema estrela é o modelo padrão para esquematizar a arquitetura de data marts dimensionais e data warehouses. É composto por uma ou mais tabelas de fatos que indexam um uma quantidade variável de dimensões.

Tabelas de fatos se relacionam com todas as dimensões, sendo o centro de convergência para todas as relações no esquema, e todas as dimensões não se relacionam entre si, mas conectam-se necessariamente aos fatos, o que forma um modelo organizacional semelhante a uma estrela, como mostrado na figura abaixo.

Figura 16 - Esquema Estrela (Star Schema)



Fonte: [geeksforgeeks.org](https://www.geeksforgeeks.org/)

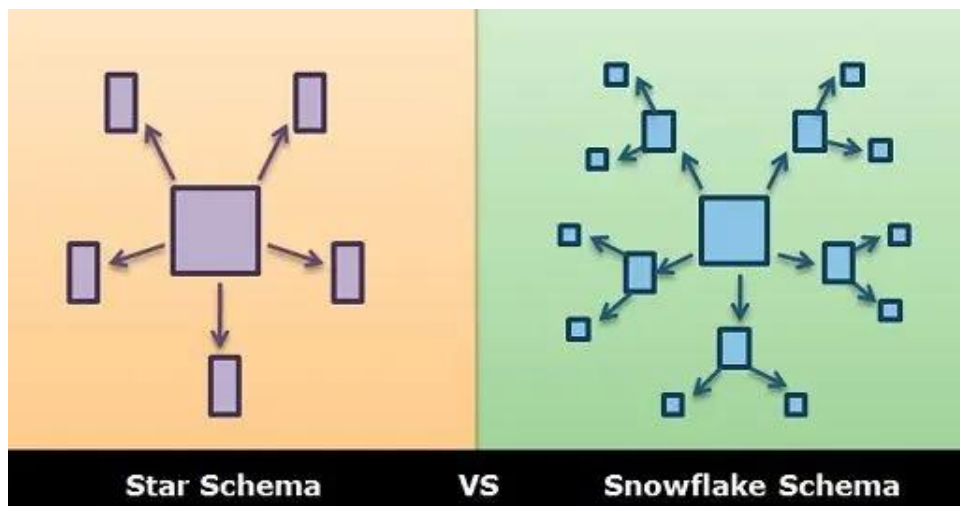
Na figura 16 acima, a coluna de vendas é a coluna fato, que abrange uma ocorrência concreta no mundo real e cada dimensão se relaciona a ela, cada uma mantendo registro de valores específicos a um fim, como os dados do cliente, do vendedor, da data e do produto. Esta organização visual é o que motiva a nomenclatura do modelo de star schema(SSM).

No SSM, todos os dados de processos de um negócio estão distribuídos em uma ou várias tabelas de fato, se forem quantitativos e em tabelas de dimensões, se forem características descritivas das partes envolvidas no negócio referenciado pelo fato. Em suma, a tabela fato armazena valores quantitativos de do domínio de negócio e as dimensões armazenam dados qualitativos que descrevem este domínio e o seus fatos.

O modelo de snow flake(SFM) são materializados quando as dimensões de um SSM possuem um nível de detalhamento e estrutura elevados, com diversos níveis de relacionamento entre eles incluindo relações de hierarquia entre as dimensões, que podem se relacionar de forma unidirecional de forma hierárquica em uma relação de super e sub dimensão.

Desta forma, o snowflake é um caso especial e uma implementação mais robusta de um star schema que considera nivelamento elevado de relacionamentos e hierarquia de dimensões. Na imagem a seguir pode-se ver a comparação visual dos dois esquemas.

Figura 17 - Esquema flocos de neve(Snowflake Schema)



Fonte: hevodata.com

2.13 CUBOS OLAP

Os cubos cruzam as diversas dimensões de dados existentes em um banco de dados DW/DM e os integram em um conjunto organizado e coeso, que por isto retratado como um modelo de dados multidimensional(cubo). Esta abordagem permite com que seja feito a mescla da dimensão com o fato e a junção de todos eles conforme o tipo de análise aplicado, possibilitando focalização ou generalização da análise. Dimensão é um aspecto da informação que classifica um dado em um agrupamento bem definido, como endereço, dados pessoais ou informações de um produto. Os cubos de um banco de dados baseado em On-line Analytical Processing caracterizam-se por formato de armazenagem dos dados estruturados em várias dimensões intercambiáveis e correlacionáveis. Os cubos OLAP, atualmente, dividem-se nas estruturas multidimensional e tabular.

2.13.1 Multidimensional

Os cubos Olap Multidimensionais ou MOLAP diferem dos bancos de dados relacionais tradicionais ao tratar cada dimensão em separado. As dimensões são tratadas como uma

entidade atômica e podem ser cruzadas para extrair informações analíticas. Cada entidade de um banco de dados relacional pode ser subdividida em várias dimensões contendo as informações normalizadas e um fato, que é o centralizador de correlação entre elas.

2.13.2 Tabular

Os cubos tabulares são uma implementação limitada dos antigos multidimensionais que utilizam a linguagem de expressões para análise de dados (do inglês, Data Analysis Expressions ou DAX). Armazenam dados em formato colunar dentro de tabelas, que são divididas em duas dimensões, a de fato e a de medidas. Cada entidade de um banco relacional pode ser dividida nestas dimensões e podem ser cruzadas para aplicar as análises. São modelos relacionais de cubos, portanto ROLAP (relacional on-line analytical processing).

2.14 CLOUD SERVICES (SERVIÇO EM NUVENS)

A Cloud Service está cada vez mais difundido na sociedade e está em um progresso de utilização, além disso há o contexto de preservação dos tipos de objetos encontrados nos ambientes digitais, facultando o armazenamento e o acesso aos dados a longo prazo.

A expressão "*cloud services*" concerne a um vasto encadeamento de serviços proporcionados sob demanda para organizações e clientes através da Internet. Esses serviços são planejados para propiciar facilidade no acesso a aplicações e recursos, sem necessariamente de infraestrutura local. Podemos averiguar isso a partir da verificação de e-mail à colaboração em documentos, a maioria dos usuários usufruem os serviços em nuvem no decorrer das atividades laborais, estejam eles cientes disso ou não.

Os "*cloud services*" são administrados por fornecedores e provedores de serviços de computação em nuvem. Eles são fornecidos aos clientes a partir dos servidores dos provedores, por conseguinte, não há necessidade de uma organização hospedar os aplicativos em seus próprios servidores locais.

Podemos afirmar que as principais vantagens da utilização de serviços em nuvem são:

A escalabilidade: Não se faz necessário uma organização investir em seus próprios recursos ou designar uma equipe de TI extra para administrar o serviço, em razão de que o provedor de serviços em nuvem fornece toda a infraestrutura e softwares necessários.

Economia: A grande maioria dos serviços em nuvem são ofertados como subscrição mensal ou anual, nesse caso não há necessidade de custear licenças de software localmente. Isso gera uma grande economia, pois ela poderá acessar os serviços de software, armazenamento e outros serviços sem há necessidade de investir em uma infraestrutura subjacente ou lidar com manutenção e atualizações.

Flexibilidade: As organizações podem adquirir serviços sob demanda, conforme a necessidade. No momento que não houver mais necessidade de uma determinada aplicação ou plataforma, a empresa pode simplesmente cancelar a assinatura ou encerrar o serviço.

Sobre a decisão de como aproveitar os serviços de nuvem, as organizações podem eleger que tipo de ambiente opera melhor para o seu negócio: nuvem pública, nuvem privada ou uma combinação de ambas.

Privado: Os serviços de nuvens privadas são construídos somente para uma única organização. Diferentemente de um data center privado virtual, a infraestrutura utilizada pertence a empresa, à vista disso, ele tem controle total sobre como as aplicações são implementadas na nuvem. Uma nuvem privada geralmente é criada sobre um data center privado.

Público: Os serviços de nuvens públicas são executados por terceiros. Ficando misturado as aplicações de diversos usuários nos sistemas de armazenamento, isso pode transparecer uma ação ineficiente a princípio. Mas isso se justifica, visto que, se a implementação de uma nuvem pública julga demandas essenciais, como por exemplo, o desempenho e segurança, logo, a existência de outras aplicações sendo executadas na mesma nuvem continua transparente tanto para os prestadores de serviços como para os usuários.

Comunidade: Neste modelo a infraestrutura do serviço de nuvem é compartilhado por várias organizações e suporta uma comunidade específica que compartilha as responsabilidades. Esse modelo pode ser gerenciado por organizações ou por um terceiro e pode existir localmente ou remotamente.

Híbrido: Nos serviços de nuvens híbridas há uma conjuntura dos moldes de nuvens públicas e privadas. Permitindo assim que um serviço nuvem privado possa ter seus recursos expandidos a partir de uma retenção de recursos em uma nuvem pública. Há uma vantagem nesse atributo que é reter os níveis de serviço mesmo que haja oscilações esporádicas na necessidade dos recursos.

2.14.1 SaaS

Software como serviço ou SaaS. Essa extenso conjunto envolve diversos serviços, como armazenamento e backup de arquivos, e-mail baseado na Web e ferramentas de gerenciamento de projetos.

Nessas aplicações, os usuários podem acessar, compartilhar, armazenar e proteger informações na “nuvem”.

2.14.2 IaaS

Infraestrutura como serviço, ou IaaS, dispõe a infraestrutura de que muitos provedores de serviços em nuvem precisam para gerenciar ferramentas SaaS, todavia não querem se manter. Esse serviço atua como a estrutura completa do data center, eliminando a necessidade de instalações no local com uso intensivo de recursos.

Os provedores mantêm todos os servidores de armazenamento e hardware de rede, além disso também podem oferecer balanceamento de carga, firewalls de aplicações, entre outros.

2.14.3 PaaS

O modelo de serviço em nuvem conhecido como plataforma como serviço, ou PaaS, serve como um ambiente baseado na web onde os desenvolvedores podem construir aplicativos em nuvem. PaaS fornece um banco de dados, sistema operacional e linguagem de programação que as organizações podem usar para desenvolver software baseado em nuvem, sem ter que manter os elementos subjacentes.

2.14.4 DaaS

“Development as a Service” ou Desenvolvimento como Serviço: Neste serviço as ferramentas de desenvolvimento tomam forma no cloud computing como ferramentas compartilhadas, ferramentas de desenvolvimento web-based e serviços baseados em mashup.

2.14.5 CaaS

“Communication as a Service” ou Comunicação como Serviço: Neste serviço é usado uma solução de Comunicação Unificada hospedada em Data Center do provedor ou fabricante.

2.14.6 EaaS

“Everything as a Service” ou Tudo como Serviço: Este serviço é quando se utiliza tudo, infraestrutura, plataformas, software, suporte, enfim, o que envolve T.I.C. (Tecnologia da Informação e Comunicação) como um Serviço.

2.14.7 Serviços escolhidos

Nesta seção, são detalhadas as definições de cada serviço em nuvem escolhido no conjunto de ferramentas e tecnologias para implementação do trabalho.

2.14.7.1 Azure

Azure é a plataforma de cloud computing (computação em nuvem) da Microsoft, que oferece diversos serviços de construção, publicação, testes e gerenciamento de aplicativos distribuídos on-line por toda a rede de data-centers por eles gerenciadas. Fornecem vários serviços na modalidade IaaS, PaaS e SaaS e provê suporte as mais variadas linguagens de programação, ferramentas e frameworks de desenvolvimento, incluindo os da própria Microsoft e de outras companhias terceiras.

2.14.7.2 Azure Data lake Gen1

O Azure Data Lake Storage Gen1 é um repositório de larga escala extensível a todo negócio, projetado para cargas de trabalho em big data analytics. Ele permite captar dados de diferentes tamanhos, tipos e velocidades de ingestão em um único lugar para análise exploratória e operacional de dados.

O ADLS G1 pode ser acessado do Hadoop, disponível com o serviço de cluster HDInsight, utilizando REST APIs compatíveis com WebHDFS. Foi projetado para permitir análise de dados sobre os dados armazenados e é ajustado para alta performance para os cenários de análise de dados. O serviço inclui todas as capacidades de janela empresarial: segurança, gerenciabilidade, escalabilidade, confiabilidade e disponibilidade.

2.14.7.3 Azure SQL Database

Azure SQL Database é uma plataforma como serviço (platform as a service ou PaaS) de ferramenta de banco de dados totalmente gerenciada que suporta praticamente todas as funções administrativas de bancos de dados, tais como melhoria de versão(Upgrading), atualização de versão(patching), cópias de segurança(backups) e monitoria de sistema sem envolvimento do usuário. O serviço SQL DB sempre roda com base na versão mais estável da ferramenta de bancos de dados do Microsoft SQL Server e seu SO despachado possui disponibilidade de 99.9%. As capacidades de um PaaS embutidas no serviço SQL DB permitem ao desenvolvedor em administração e otimização de bancos de dados específicos para domínios específicos de criticidade para um determinado negócio.

2.14.7.4 Azure Data Factory

Azure Data Factory é o serviço ETL na nuvem do Azure para integração e transformação de dados sem servidor em expansão. Ele oferece uma interface do usuário livre de código para criação intuitiva e gerenciamento e monitoramento em painel único. Também é possível migrar pacotes SSIS existentes por lift-and-shift para o Azure e executá-los com total compatibilidade no ADF (Azure Data Factory). O Azure-SSIS Integration Runtime oferece um serviço totalmente gerenciado, de modo que não é necessário se preocupar com o gerenciamento da infraestrutura.

2.14.7.5 Azure Analysis Services

O Azure analysis services é uma plataforma como serviço (platform as a service ou PaaS) totalmente gerenciada, que oferece a janela empresarial modelos de dados em nuvem. Utiliza-se junções e alterações para combinação de dados de múltiplas fontes, definir métricas e assegurar os dados em uma estrutura tabular semântica única e confiável. O modelo construído fornece uma maneira mais fácil e rápida de performar análises de dados “Ad Hoc” com emprego de ferramentas como PowerBI e Excel.

2.15 FORMATOS DE DOCUMENTOS DE TEXTO

Documentos textuais são as principais ferramentas para armazenar dados de forma não estruturada ou semiestruturada, pois permitem a troca de armazenamento destes dados entre sistemas. São utilizados em processos de ingestão. Há vários formatos com diferentes estruturas, neste trabalho são utilizados o JSON na documentação do ADF para organizar e codificar a orquestração de movimentação de dados e o CSV, que é o formato dos arquivos brutos de fonte de dados, no qual se encontra a base de micro dados do ENEM e os conjuntos de dados para preparação e transformados.

JSON: Os dados JSON (JavaScript Object Notation) são representados como pares chave-valor em um formato semiestruturado. O JSON costuma ser comparado com o XML, pois ambos podem armazenar dados em formato hierárquico, com os dados filho representados embutidos com seu pai. Ambos são auto descritivos e legíveis por humanos, mas os documentos JSON tendem a ser muito menores, resultando em seu uso popular na troca de dados online, especialmente com o advento de serviços Web baseados em REST.

CSV: Os arquivos CSV (valores separados por vírgula) são geralmente usados para troca de dados de tabela entre sistemas em texto sem formatação. Eles normalmente contêm uma linha de cabeçalho que fornece nomes de coluna para os dados, mas de outra forma, são considerados semiestruturados. Isso é devido ao fato de que os CSVs não podem representar dados hierárquicos ou relacionais naturalmente. As relações de dados costumam ser manipuladas com vários arquivos CSV, em que as chaves estrangeiras são armazenadas em colunas de um ou mais arquivos, mas as relações entre esses arquivos não são expressas no próprio formato. Arquivos no formato CSV podem usar outros delimitadores além de vírgulas, como tabulações ou espaços.

2.16 VISUALIZADORES DE DADOS

A visualização de dados consiste em representações gráficas de informações e dados. Usando elementos de visualização (como tabelas, gráficos e mapas), a visualização de dados é uma maneira conveniente de visualizar e compreender anomalias, tendências e padrões nos dados. As ferramentas e técnicas de visualização de dados são essenciais para analisar grandes quantidades de informações e tomar decisões baseadas em dados.

A visualização de dados é uma forma de arte visual que atrai a atenção e mantém o foco nas informações veiculadas. Quando se olha para um gráfico, pode se visualizar imediatamente tendências e anomalias. É internalizado rapidamente tudo o que foi visto.

Com o advento da "era do big data", a visualização é uma ferramenta mui útil que pode ser usada para interpretar e entender os trilhões de linhas de dados que podem ser gerados todos os dias. A visualização de dados ajuda a contar a história, compilando os dados em um formato mais compreensível, destacando tendências e anomalias.

Conquanto, isso não se limita a projetar estilos gráficos para torná-los mais bonitos ou preencher informações em gráficos. Para criar uma visualização de dados eficaz, a sensibilidade é necessária para equilibrar a aparência e funcionalidade. Um gráfico extremamente simples pode ser monótono demais para despertar o interesse, assim como pode transmitir uma ideia influente.

Tipos de visualização de dados:

- Gráficos
- Tabelas
- Diagramas
- Mapas
- Infográficos
- Painéis
- Gráfico de área
- Gráfico de barras
- Gráfico de caixa
- Nuvem de bolhas
- Gráfico de marcador
- Cartogramas
- Exibição de círculos
- Mapa de distribuição de pontos
- Gráfico de Gantt
- Mapa de variações
- Tabela de destaque
- Histograma
- Matriz
- Rede
- Área polar
- Árvore radial

- Gráfico de dispersão (2D ou 3D)
- Gráfico de fluxo
- Tabelas de texto
- Linha do tempo
- Mapa de árvore
- Gráfico de segmentos
- Nuvem de palavras

2.16.1 Visualizador escolhido: Power BI

O visualizador de dados Power BI é uma junção de aplicações, software e conectores que convertem fontes de dados não relacionadas em informações coesas. A origem fonte pode advir de um arquivo em Excel ou de um acervo de data warehouses híbridos locais ou baseado em nuvem.

O Power BI é composto por três elementos que podem funcionar juntos:

- Um aplicativo de desktop do Windows chamado Power BI Desktop.
- O serviço SaaS online (software como serviço) é chamado de serviço Power BI.
- Aplicativo móvel Power BI para dispositivos Windows, iOS e Android.

Esses elementos supracitados são projetados para permitir que se crie, compartilhe e se use insights de negócios de maneira eficaz.

- Ademais, o Power BI também tem dois outros elementos: Gerador de relatórios do Power BI, usado para criar relatórios paginados para compartilhar no serviço Power BI.
- Depois que o servidor de relatório do Power BI é criado na área de trabalho do Power BI, pode-se publicar o servidor de relatório local do Power BI nele.

O fluxo de trabalho comum no Power BI começa com a conexão a fontes de dados no Power BI Desktop e a criação de relatórios. Em seguida, pode-se publicar o relatório do Power BI Desktop para o serviço do Power BI e compartilhá-lo para que os usuários de negócios no serviço do Power BI e dispositivos móveis possam visualizar e interagir com o relatório. Esse fluxo de trabalho é muito comum e mostra como os três elementos principais do Power BI se integram.

No serviço Power BI, pode-se usar a ferramenta de pipeline de implantação para testar o conteúdo antes de iniciar para o usuário. A ferramenta de pipeline de implantação tem como

objetivo, auxiliar na implantação de relatórios, painéis, conjuntos de dados e relatórios paginados.

2.16.2 Power Query/M language (Marcação)

O Power Query é um suplemento do Excel desenvolvido pela Microsoft e seu objetivo é facilitar o carregamento de dados no Excel a partir de fontes de dados externas. Faz parte do conjunto de ferramentas Power BI e, como tal, é apenas uma ferramenta entre muitas que se pode usar ao desenvolver o que são chamadas de "soluções de Business Intelligence de autoatendimento". O objetivo de uma solução de Business Intelligence (geralmente abreviado para "BI") é tornar as informações de negócios acessíveis às pessoas para que possam usá-las para tomar decisões informadas sobre como fazer seu trabalho.

Todo o processo é referido como "self-service" porque as pessoas que querem usar esses dados, os analistas, os contadores, os gerentes e assim por diante, também são as pessoas que estão construindo os relatórios.

Dado que o Power Query é apenas uma parte do pacote Power BI, é importante observar todos os componentes do Power BI para que se possa colocar o Power Query em contexto, entender o que ele faz, entender quando se deve usá-lo e quando outra ferramenta é mais apropriada.

O Power Query é o primeiro componente do conjunto do Power BI a ser usado ao criar uma solução de BI. Isso porque a primeira etapa na criação de qualquer tipo de relatório ou painel é obter os dados de origem. O Power Query permite que a conexão a uma ampla variedade de fontes de dados diferentes. É possível extrair dados com rapidez e facilidade e definir uma série de etapas repetíveis para limpar, filtrar e transformar os dados antes que sejam carregados. O Power Query oferece a opção de carregar dados diretamente.

2.17 LINGUAGEM PARA ANÁLISE DE DADOS

No que tange a análise de dados, existem algumas linguagens que são empregadas para diferentes fins, seja análises de medidas, exploratórias ou de padrões estatísticos. Neste tópico são listadas cada uma delas que foram utilizadas neste trabalho.

2.17.1 DAX

Expressões de análise de dados (DAX) é uma linguagem de expressão de fórmula usada no Analysis Services no Excel, Power BI e Power Pivot. As fórmulas DAX contêm funções, operadores e valores para realizar cálculos e consultas avançadas nos dados das tabelas e colunas relacionadas do modelo de dados tabular.

As fórmulas DAX são usadas para:

- Métricas
- Colunas calculadas
- Tabelas calculadas
- Segurança em nível de linha.

O DAX oferece suporte a vários tipos de dados, como "inteiro" (somente inteiro), "decimal" (somente parte de um inteiro), "moeda" (combinação de inteiro e decimal), "Booleano", (resultado correto ou incorreto), "string" (que é texto) e "binário" (que é um objeto, como um arquivo ou imagem).

2.17.2 R language and environment

R é uma linguagem e um ambiente de desenvolvimento para visuais gráficos e computação estatística, que se assemelha a “linguagem S” desenvolvida na Bell Laboratories por John Chambers e seus parceiros. A R é considerada uma implementação diferencial da S e apesar das diferenças, a maior parte dos seus códigos são compatíveis.

A linguagem R provê uma série variada de técnicas estatísticas de visualização gráfica, que são altamente extensíveis. Enquanto a linguagem S é conhecida como uma ferramenta frequentemente utilizada para pesquisa com metodologia estatística, o R estabelece uma rota de código aberto (open source) para esta prática.

O ambiente de programação R está disponível como um software livre através dos termos da General Public License da Free Software Foundation na forma de código fonte. É compilado e executado em uma variedade de plataformas UNIX e sistemas operacionais similares (como FreeBSD e Linux), Windows e MacOS.

Todo o ambiente compreende um espaço integrado de dispositivos de software para manipulação de dados, cálculos e disposições gráficas. Isto inclui:

- Uma ferramenta eficiente para manuseio e armazenamento de dados
- Um ambiente de operadores para cálculos com vetores, especialmente matrizes

- Uma coleção vasta, coerente e integrada de ferramentas intermediárias para análise de dados
- Uma linguagem de programação bem desenvolvida, simples e efetiva que inclui estruturas condicionais e de repetição, funções recursivas definidas pelo desenvolvedor e espaços de IO

2.17.2.1 CRAN

CRAN é uma sigla em inglês para Comprehensive R Archive Network (Rede de arquivos compreensiva do R). A CRAN é uma rede FTP e de webserveres espalhada por todo o mundo que armazena versões de código e documentação da linguagem R idênticas e atualizadas. Trata-se de uma grade computacional provisora de bibliotecas e documentações R em escala internacional com replicação de contingência e alta disponibilidade, que possibilita acesso rápido a estes recursos através da conexão com espelhos (mirrors) de maior proximidade geográfica.

2.17.3 SQL

SQL significa Structured Query Language. Formalmente, é pronunciado "Ess qui el". Como o nome sugere, a linguagem SQL é uma linguagem de consulta usada para interagir com bancos de dados relacionais. A linguagem SQL é usada para executar funções, como inserir dados em um banco de dados, recuperar dados, atualizar dados, excluir dados e outras operações semelhantes.

SQL é uma linguagem declarativa com sintaxe relativamente simples e concentra-se em bancos de dados relacionais. Pode ser aprendida por profissionais que não são necessariamente desenvolvedores, mas usualmente utilizam bancos de dados.

O SQL surgiu para padronizar a maneira como os profissionais de TI executam comandos em seu SGBD (Sistema Gerenciador de Banco de Dados). Bancos de dados muito populares (como banco de dados Oracle, MySQL, PostgreSQL e Microsoft SQL Server) são alguns especialistas em SQL. Sumariamente, podemos dizer que o SQL é direcionado a desenvolvedores e profissionais que têm um relacionamento direto com o banco de dados para manipular ou visualizar os dados com mais facilidade.

A padronização SQL não apenas mantém seus comandos em um bloco de operação, mas também os mantém em vários blocos de operação. Essas especificações são chamadas de subconjuntos.

Linguagem de processamento de dados DML ou linguagem de processamento de dados é uma linguagem de processamento de dados. Este subconjunto foi projetado para alterar os dados na tabela, como inserir, excluir e atualizar dados. Seus comandos mais comumente usados são inserir, excluir e selecionar.

A Linguagem de Definição de Dados ou DDL (Data Definition Language), é a uma linguagem de definição de dados. Esta linguagem está mais associada ao próprio banco de dados do que os dados que ele armazena. Os comandos mais utilizados são create e drop.

A Linguagem de Controle de Dados ou DCL (Data Control Language), é a linguagem de controle de dados. Esta linguagem também está mais associada à manutenção do banco de dados do que aos dados registrados por ele. Este é um subconjunto importantíssimo, pois é ele quem define permissões, bloqueios e restrições de usuários. O principal comando é o grant, que fornece acesso ou privilégios para usuários a diferentes tabelas.

A Linguagem de Transação de Dados ou DTL (Data Transaction Language), é a linguagem de alteração de dados dentro de uma tabela. Ela existe porque, antes de modificar algum dado, é preciso também autorizar que elas sejam salvas. Esse subconjunto serve para de fato publicar estas alterações através de comandos como o commit.

A Linguagem de Consulta de Dados ou DQL (Data Query Language), é a mais conhecida para quem não precisa necessariamente conhecer o funcionamento de um banco de dados, mas somente consultar suas informações. O principal comando deste subconjunto é o select.

2.18 FERRAMENTAS DE DESENVOLVIMENTO

As ferramentas utilizadas para desenvolvimento dos processos operacionais neste trabalho resumem-se em IDEs para aplicação da análise de dados OLAP e do data mining da base transformada, além de SGBDs para modelagem, pré-processamento e transformação de dados e um storage explorer para gerenciar o armazenamento dos arquivos brutos e transformados no modelo e gerenciamento do data lake. Serão listados a seguir cada um destes, contextualizando seu conceito.

2.18.1 Storage Explorer

O Azure storage explorer é um software on-premises de gerenciamento de armazenamento em nuvem, utilizado para realizar uploads e downloads de entidades como Azure blobs, arquivos, filas, e tabelas e gerenciar storages e serviços de bancos de dados da azure tais como Azure Cosmos DB e Azure Data Lake Storage. Habilita a gerência e configuração das regras de compartilhamento de recursos de origem cruzada (do inglês, Cross-Origin Resource Sharing ou CORS).

2.18.2 IDEs e DBMS

Os ambientes integrados de desenvolvimento (do inglês, Integrated development environment) são ferramentas integradas em um programa que auxilia os desenvolvedores a consolidar aspectos e passos distintos na escrita de um programa. São ferramentas que incrementam a produtividade da programação por reunir etapas comuns na escrita de softwares em uma aplicação única, são elas a edição do código fonte, a compilação e construção de executáveis e a depuração(debugging).

2.18.2.1 Visual Studio

A IDE do Visual Studio é um ambiente de desenvolvimento integrado, sendo assim, um painel de inicialização criativo que pode ser usado para editar, depurar e compilar o código e, em seguida, publicar um aplicativo.

Além do editor e do depurador padrão oferecidos pela maioria dos IDEs, o Visual Studio adiciona compiladores, ferramentas de preenchimento de código, designers gráficos e muitos outros recursos para facilitar o processo de desenvolvimento de software.

Projetos de modelo de tabela e multidimensionais são gerados utilizando os modelos de projeto no Visual Studio com as extensões de projetos de Analysis Services (VSIX). Os modelos de projeto fornecem designers de modelo e assistentes para a criação de objetos de modelo de dados que compõem uma solução de Analysis Services. As extensões de projetos do Analysis Services têm suporte em todas as edições do Visual Studio, incluindo a edição gratuita da Comunidade.

2.18.2.2 SQL Server

O Sql Server é um banco de dados relacional criado, mantido e distribuído pela Microsoft. Suporta o padrão de linguagem ANSI SQL, o modelo estrutural padronizado da linguagem e também conta com a interpretação proprietária de SQL conhecida como Transact-SQL (ou T-SQL). A linguagem T-SQL dispõe peculiaridades próprias como modelos para declaração de variável, tratamento de exceções, stored procedures entre outros. O principal SGBD utilizado para gerenciar e manipular bancos de dados MSSQL (Microsoft Sql Server) é o SQL Server Management Studio (SSMS).

2.18.2.3 R Studio

R Studio é um Ambiente de Desenvolvimento Integrado (IDE) para a linguagem de programação R. RStudio é um projeto de código aberto destinado a combinar os vários componentes do R (console, edição de código-fonte, gráficos, histórico, ajuda, etc.) em uma bancada de trabalho contínua e produtiva. Ele foi projetado para facilitar a curva de aprendizado para novos usuários da linguagem de programação R, bem como fornece ferramentas de alta produtividade para usuários mais avançados. O RStudio pode ser implementado como um servidor para permitir o acesso à web para sessões R em execução em sistemas remotos.

Os principais recursos que o R Studio dispõe são:

- Os principais componentes de um IDE são todos muito bem integrados em um layout de quatro painéis que inclui um console para sessões R interativas, um editor de código-fonte com guias para organizar os arquivos de um projeto e painéis com guias para organizar componentes menos centrais.
- O editor de código-fonte pode ser usado de uma forma simples, ele também é rico em recursos, possui excelentes recursos de navegação de código e está bem integrado ao console embutido.
- O console e o editor de código-fonte estão estreitamente vinculados ao sistema de ajuda interno de R por meio do preenchimento da guia e do componente visualizador de página de ajuda.
- O recurso de projeto facilita a organização de diferentes fluxos de trabalho.
- O R Studio fornece muitas ferramentas administrativas convenientes e fáceis de usar para gerenciar pacotes, área de trabalho, arquivos e muito mais.

- A IDE estar disponível para os três principais sistemas operacionais e pode ser acessado por meio de um navegador remoto.
- Comparando o R Studio com outras plataformas, o RStudio é muito mais simples de aprender que Enacs, mais fácil de configurar e instalar que Eclipse e StateT, tem um editor melhor que JGR, é mais organizado que Sciviews e, ao contrário de Notepad ++ e R6u, está disponível em mais plataformas do que apenas Windows.

O programa R Studio pode ser executado no desktop ou por meio de um navegador. A versão desktop está disponível para plataformas Windows, Mac OS X e Linux e se comporta de maneira semelhante em todas as plataformas, com pequenas diferenças para os atalhos de teclado.

2.18.3 DBMS(SGBD)

Os SGBDs ou sistemas de gerenciamento de bancos de dados (do inglês, Data Base Managment Systems) é uma coleção de diversos programas que permite o desenvolvedor de bancos de dados construir, manipular e gerenciar bancos de dados com as mais diversas finalidades. Cada SGBD tem suas ferramentas integradas e diversas, porém todos eles devem prover tecnologias para atender as seguintes demandas:

- Controle de Redundâncias - Armazenamento em um único local evitando duplicações descontroladas;
- Compartilhamento de Dados;
- Controle de Acesso;
- Interfaceamento - Disponibilizar versões gráficas e não somente modo texto;
- Esquematização - Tornar compreensível as relações entre tabelas;
- Controle de Integridade;
- Cópias de Segurança

2.18.3.1 Azure Data Studio

O Azure Data Studio é uma ferramenta de banco de dados multiplataforma para profissionais de dados que usam plataformas de dados on-premises e em nuvem no Windows, macOS e Linux.

O Azure Data Studio oferece uma experiência moderna de edição de código com ‘IntelliSense’, trechos de código, integração de controle de origem e um terminal integrado. Ele é projetado com o usuário de plataformas de dados em mente, com o gráfico de conjuntos de resultados de consulta embutido e dashboards personalizáveis.

2.18.3.2 SQL Server Management Studio(SSMS)

O SSMS é um ambiente integrado para gerenciamento de qualquer infraestrutura SQL, desde o SQL Server(on-premises) até o Azure SQL Database(on-cloud). Provê ferramentas para configurar, monitorar instâncias de SQL Server e seus bancos de dados. Utilizado para construir, implantar e aprimorar componentes no nível de dados e desenvolver queries de consulta e scripts de bancos de dados.

2.19 Git e Github

Git é um projeto de código aberto, criado por Linus Torvalds em 2005, e ativamente mantido por ele de forma constante até os dias de hoje. É fortemente adotado pela indústria de desenvolvimento de softwares e aplicações para controle e versionamento dos seus códigos-fonte. Opera em uma arquitetura distribuída, por isto é um exemplo de um sistema distribuído para controle de versão. As vantagens do git para versionar códigos de projetos e produtos, é a garantia de performance, com compressões de arquivos e estrutura de árvores para identificar alterações, ao invés de referenciá-las por nome e tipo de arquivo. Também dispõe uma forte garantia de segurança, pela implementação de criptografia ponta a ponta no padrão SHA1 e chaves SSH para encriptar repositórios remotos. Também dispõe uma vasta flexibilidade, por permitir desenvolvimento e fluxo de projetos não lineares, através da subdivisão da árvore de alterações em ramos (do inglês, branches), para diferenciar as áreas de atuação em um repositório, e também permite bifurcação (do inglês, fork) que habilita o compartilhamento de um projeto aberto para qualquer pessoa reutilizar.

O Github é um website SaaS que implementa o protocolo Git para controle de versão de códigos. É distribuído em nuvem e de acesso aberto, permite criação de repositórios públicos e privados, além de organização de projetos associados a equipes delimitadas. Fornece uma camada de hospedagem de uma computação em grades para versionamento git a ser utilizado

por organizações e desenvolvedores individuais, para diversos fins. Possui uma parceria estratégica com a International Business Machine (IBM), e foi fundado em 29 de janeiro de 2008 por Chris Wanstrath, PJ Hyett, e Tom Preston-Werne, nos seus escritórios em São Francisco, California (EUA).

3. PROJETO DE BI MODERNO COM MINERAÇÃO E ANÁLISE DE DADOS PARA ESTUDO DO ENEM 2019

Neste capítulo, constam a descrição e evidências de todos os processos no desenvolvimento da ferramenta, na pesquisa de campo e na operação das tecnologias e técnicas aplicadas de business intelligence e data science.

3.1 DESCRIÇÃO DO PROCESSO OPERACIONAL

O presente trabalho utilizou técnicas de engenharia de dados para movimentação de dados entre diferentes fontes para serem carregados em destinos específicos para preparação(stagging), transformação, treinamento e consulta.

Na etapa de engenharia, foi realizado a estruturação e alimentação de um armazenamento datalake para subir os dados brutos, que seriam transportados e adaptados para os bancos de dados de preparo e transformação, com as compatibilizações devidas. Para transportar os dados entre diferentes serviços, utilizou-se um framework PaaS da Azure, o Data Factory, através de pipelines de cópias de dados.

Através deste serviço, foi possível orquestrar os serviços de servidores vinculados de fonte de origem e coletor destino e a execução dos processos de movimentação de dados entre eles, o que permitiu transpor os dados dos repositórios data lake para os bancos SQL Database, um PaaS azure utilizado para homologar o modelo e o banco on-premisses SQL Server 2019 express, o banco para a produção do modelo de dados em star schema no padrão de data mart relacional.

No banco PaaS SQL DB foi esquematizado um modelo padrão e na instância SQL Server on-premisses este modelo foi maturado e carregado com a ingestão dos dados provenientes da orquestração via ADF, que aplicou a compatibilidade e transposição dos tipos entre arquivo CSV bruto original da fonte e os tipos de dados no modelo destinado. Este foi o processo de Extract Load and Transform, para a produção do modelo de dados para aplicação das análises. O ADF utiliza o padrão de documentos JSON para realizar a mensageria entre origem e destino e comunicar as transformações nos dados.

O modelo de RDM criado no banco chamado EnemRDM foi construído com todos os dados da base, contendo a tabela fato chamada de EnemFactTable, que contém todas as notas das provas e redação, e as tabelas de dimensão que são a DimensionTableInfoPessoa, destinada aos dados pessoais do participante, DimensionTableInfoProva, que contém os dados a respeito

da prova realizada, DimensionTableSTEM, que armazena dados acerca da situação de conclusão de ensino médio, a tabela DimensionTableQSE, que guarda os dados referentes as respostas dos questionários socioeconômicos e a tabela DimensionTableInclusao, que é o destino de todas as informações a respeito de necessidades de atendimento especializado ou específico, bem como dados sobre uso de recursos adaptados.

No banco de dados, as colunas que não teriam utilidade foram descartadas na fase de seleção de dados, como as correspondentes a gabarito e a taxa de respostas da prova, utilizadas apenas para medição e controle dos cálculos de nota, portanto sendo irrelevantes devido a existência da nota final de cada área de conhecimento. As tabelas de dimensão e a fato foram varridas com uma consulta total(fullscan) para gerar um arquivo CSV contendo suas informações para que fossem carregadas para o data lake na área de staging da análise de dados e serviram para alimentar o modelo de cubo OLAP. O processo ELT analítico utilizou o carregamento direto das tabelas do data lake.

A tabela de inclusão não será aproveitada para a mineração de dados, pois não é o foco do trabalho, porém será utilizada na análise de dados ad-hoc com a utilização do cubo olap para testes de hipóteses contra esta base transformada, para extrair informações quantitativas em um levantamento estatístico. Para fins de data mining, todas as demais tabelas serão utilizadas, portanto as outras que não a de inclusão foram mescladas e agrupadas em uma tabela unificada chamada JoinFactDim_InfoProva_InfoPessoa_QSE_STEM. Esta tabela foi pré-processada, com a eliminação de colunas irrelevantes para mineração como os códigos de localidade e da escola, que configura a etapa de limpeza, bem como a discretização dos valores contínuos das notas em classificações discretas e o a eliminação de redundâncias com alteração de valores repetitivos com aplicação de identificadores, que tem como base um dicionário de alterações.

A transformação se deu pela alteração do tipo das variáveis de nota para números reais e depois com a criação de colunas condicionais para criar os grupos discretos com base em faixas de nota específicas, remoção de valores nulos pela identificação de cada um deles e colunas condicionais para agrupamento de valores discretos como a idade do participante. Na etapa da mineração, também foi realizado a remoção dos cabeçalhos, para não ser contabilizado pelo.

Após a modelagem e aplicação dos processos de ELT e ETL, aplicou-se a análise de dados ad-hoc contra o cubo Olap e extração da coluna pré-processada e transformada do DM para um CSV com todos os dados e foi aplicado o algoritmo de mineração para descobertas de

associação Apriori contra este arquivo. Através dele, foi possível definir associações probabilísticas entre os conjuntos de dados.

Os resultados da análise estatística de dados foram consumidos pela ferramenta de self-service BI Microsoft Power Bi, para onde o cubo foi carregado por um pivot de consulta via power query. Já os resultados do Apriori foram carregados em um dataframe no ambiente RStudio com exibição das regras em um dataframe e visualização gráfica com elementos da biblioteca RColorBrewer, para gerar os itens mais frequentes.

3.1.1 Gerência do projeto

Foi utilizada a ferramenta Planner da microsoft, disponível em tasks.microsoft.com, para controle e gerenciamento de todas as fases e ciclo de vida deste trabalho. Nele, é possível ajustar os deadlines para prazos de entrega, níveis de urgência, atribuição de agentes e definição de início e fim.

A estrutura é composta por buckets que englobam atividades em um agrupamento e cada atividade é um ticket. Os buckets criados foram o de Reunião, para controlar as reuniões com a orientadora, Parte textual, para organizar os prazos de entrega da parte escrita do documento do projeto, Questionários, para definir as metas de iniciação e coleta dos dados de cada questionário realizado e Ferramentas, para controlar os cronogramas de desenvolvimento de cada fase da ferramenta.

Para a execução dos questionários de entrevista com os profissionais da área e o feedback de avaliação dos resultados com os especialistas, foi utilizado a plataforma Microsoft Forms, que é um SaaS da Microsoft, disponível no link forms.office.com. Foram criados formulários abertos a público para quem detivesse o link gerado.

Tanto o Planner como o Forms são softwares como serviço que integram a plataforma office 365 da Microsoft, sendo acessíveis através de um e-mail corporativo ou de estudante. O Planner possui integração com a plataforma Teams, podendo ser vinculado a equipes na plataforma de teleconferência, no entanto esta funcionalidade não foi explorada para o planejamento deste trabalho.

Para a comunicação remota entre os membros da dupla e organização de pendências, foi utilizado o SaaS Discord, que é um tele mensageiro gratuito e aberto que permite a criação de salas compostas por canais de texto para discussão, canais de voz e de vídeo para apresentações.

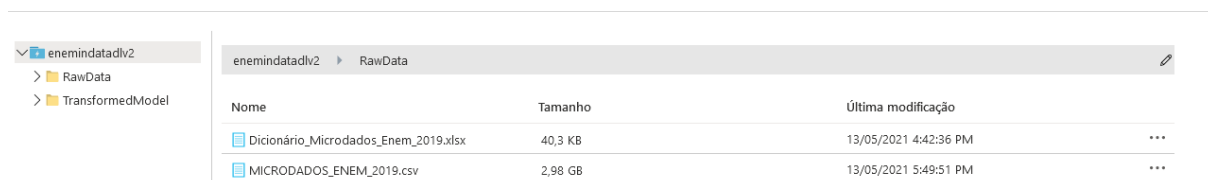
Todo o projeto, as documentações textuais geradas, os códigos de SQL e R e as imagens dos visuais e gráficos produzidos foram carregadas para um repositório git na plataforma online Github, e será constantemente alimentado enquanto houver atualizações deste trabalho nesta plataforma. O link do repositório é <https://github.com/fabricioasn/EnemInData_TCC_Unicarioca>.

A ideia desta documentação é o controle de versionamento e, para além disto, permitir uma sobrevida e registro permanente do projeto para viabilizar possíveis bifurcações a partir dele, bem como ajustes, melhorias e outras implementações incrementais que agreguem e complementem funcionalidades extras a ele. Desta forma, fica como uma fonte de consulta para projetos futuros e apontamento de caminhos práticos para orientar os rumos destes trabalhos vindouros.

3.1.2 Engenharia de dados

A engenharia de dados neste trabalho foi pautada nos moldes da arquitetura moderna de BI para Bigdata, que prevê suporte a compartilhamento e interconexão entre diferentes fontes e destinos de dados de tipos distintos em um repositório Data lake, de modo que os dados estejam disponíveis e preparados para usos diferenciados para as várias abordagens analíticas com diversas finalidades. A origem dos dados brutos para o processo de análise dos dados é única e proveniente de uma fonte estática, que é o portal de bases de dados públicas do INEP. O arquivo utilizado para aplicação das ferramentas analíticas é a base de micro dados do Enem de 2019, em formato CSV.

Figura 18 - Diretório de arquivos do Data Lake da Azure



The screenshot shows the Azure Data Lake Storage interface. On the left, a navigation pane shows the directory structure: 'enemindatadlv2' (expanded) containing 'RawData' and 'TransformedModel'. The main pane shows the 'RawData' directory with a table of files.

Nome	Tamanho	Última modificação	
Dicionário_Microdados_Enem_2019.xlsx	40,3 KB	13/05/2021 4:42:36 PM	...
MICRODADOS_ENEM_2019.csv	2,98 GB	13/05/2021 5:49:51 PM	...

Fonte: Elaborado pelo autor (2021)

O arquivo bruto foi baixado do INEP e extraído do ZIP, e tanto ele quando o dicionário dos microdados foram manualmente carregados para o ADLS Gen1 através do cliente windows ASE, devido ao tamanho de 3GB, que faz necessário o uso do explorador de arquivos dos

repositórios data lake azure, pois o serviço Web é limitado a upload de arquivos com no máximo 2GB. O objetivo de um datalake é permitir acesso a dados das mais diferentes características para fins diversificados, garantindo assim a democratização dos dados e assegurar a delimitação de visões para controlar e gerenciar a segurança do compartilhamento distribuído no repositório.

O armazenamento foi separado em uma pasta nomeada RawData, onde foi inserido o arquivo da base de dados original e a pasta TransformedRDW, para onde foram destinados os arquivos CSVs das tabelas modeladas após a seleção de dados aproveitáveis para uso na análise ad-hoc. Nesta pasta o serviço Analysis services busca e conecta-se aos conjuntos de dados utilizados na análise.

Para a ingestão dos dados no modelo criado para organização dos dados, foi utilizado o serviço ADF que abstrai clusters de Apache Spark para processamento de pipeline de dados. A modelagem segue o padrão de um data mart relacional criado em homologação para teste em um banco Azure SQL DB, utilizado para desenvolvimento do modelo e testagem dos pipelines. O recurso ADF permite estruturar um pipeline de dados de movimentação de dados em processamento Batch durante todas as etapas da arquitetura de BI e suas camadas de processos, provisionando um orquestrador de pipelines fim a fim no ciclo de vida da análise de dados e as conexões de armazenamento envolvidas.

Figura 19 - Configuração dos pipelines para o Azure SQL DB

Nome ↑↓	Tipo ↑↓	Relacionado ↑↓	Anotações ↑↓	Usar o ponto de extremidad...
 AzureDataLakeStore1	Azure Data Lake Storage Gen1	24		---
 AzureDataLakeStoreForE...	Azure Data Lake Storage Gen1	15		---
 SqlServer1  	Servidor SQL	13		---
 SqlServerDMLS	Servidor SQL	0		---

Fonte: Elaborado pelo autor (2021)

O modelo final foi definitivamente criado para produção em uma instância em máquina local de MSSQL 2019 Express, vinculado como serviço ao ADF através de uma camada de interface de redes chamada integration runtime self-hosted, que provê um proxy virtualizado em tempo de execução para transformar a instância MSSQL na máquina em um servidor para ADF. Através desta infraestrutura, foi possibilitado designar uma máquina local como destino dos pipelines de dados, onde foi realizado a modelagem do banco para a produção do conjunto de dados.

Figura 20 - Configuração do Data Factory na Azure

Nome	Tipo	Subtipo	Status	Região	Criado em
AutoResolveIntegrationRuntime	Azure	Rede Virtual...	Em execução	Resolver Automatica...	5/12/21, 8:51:42 PM
SelfHostedIntegrationRuntime1	Auto-hospe...	---	Em execução	---	5/13/21, 1:01:53 AM

Fonte: Elaborado pelo autor (2021)

Figura 21 - Integração do MSSQL da Azure

Microsoft Integration Runtime Configuration Manager

Home Settings Diagnostics Update Help

✓ Self-hosted node is connected to the cloud service

Data Factory: EnemInDataDFv2
 Integration Runtime: SelfHostedintegrationRuntime1
 Node: ALMEIDABATISTA

Stop Service

Data Source Credential ⓘ

Credential store: On-premises
 Credential status: In sync
 Last backup time: N/A

Generate Backup Import Backup

✓ Connected to the cloud service (Data Factory V2)

Fonte: Elaborado pelo autor (2021)

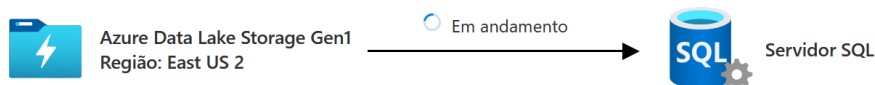
Figura 22 - Relacionamento entre as tabelas do banco de dados



Fonte: Elaborado pelo autor (2021)

Os dados foram copiados do arquivo bruto na pasta RawData do ADLS com uso do ADF para o banco local MSSQL, com a transposição dos tipos string do documento CSV para os tipos Varchar padronizados no modelo criado no banco destinado à carga dos dados. Foi utilizado, para delimitação dos campos, a opção de definir a primeira linha de cabeçalho e separação por vírgula. Também se utilizou para codificação o conjunto de caracteres UTF-7 na carga para o banco, para preservar os caracteres especiais. Cada pipeline executado teve como objetivo a carga de uma tabela, por último a carga da tabela de mescla.

Figura 23 - Conexão do Data Lake com Servidor SQL

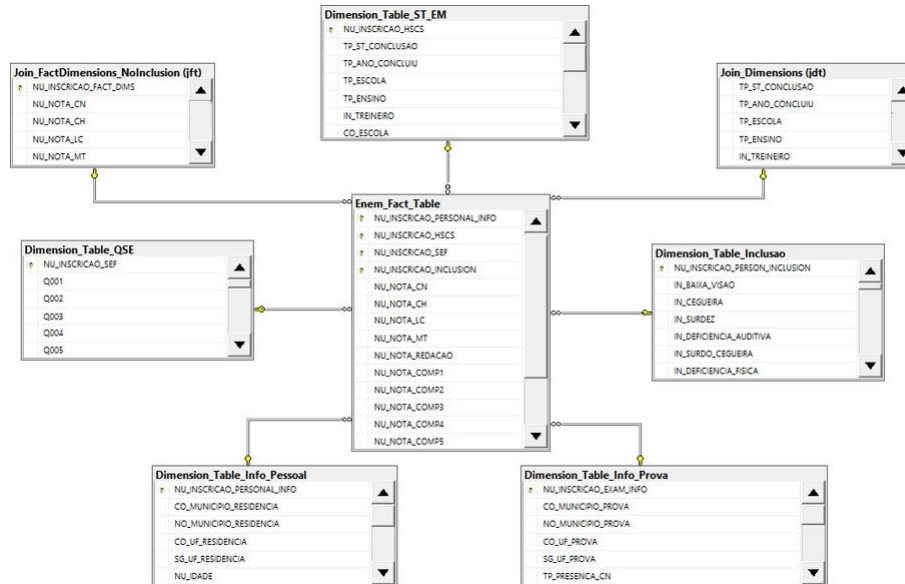


Fonte: Elaborado pelo autor (2021)

O banco EnemRDM criado no padrão star-schema possui como tabela fato os dados de notas das áreas de conhecimento e cada dimensão corresponde a informações sobre o participante sobre o exame, que são divididas em informações pessoais, informações da prova, situação do ensino médio e questionário socioeconômico. A chave de todas as dimensões é a redundância do número de inscrição, identificada para cada uma delas. Todas estas chaves são exportadas como chave estrangeira para a tabela fato, que juntas formam uma chave primária

composta nela. Dividiu-se o modelo nos esquemas “dbo”, que corresponde ao fato e as dimensões e o esquema “jft” possui a tabela de junção. As chaves primárias das tabelas de junção são chave estrangeira na “fato”. Todas as relações são 1 para 1.

Figura 24 - Relacionamento entre as tabelas do banco de dados transformado



Fonte: Elaborado pelo autor (2021)

3.1.3 Ciência de dados

O escopo da ciência de dados é delimitado pela escolha e aplicação do algoritmo Apriori na linguagem R, bem como a modelagem da base de dados e seleção de um conjunto de dados deste modelo para ser minerado. O modelo estruturado foi um data mart relacional em esquema estrela e foram aplicados os processos do KDD de coleta, com a ingestão dos dados, de pré-processamento, com o preparo do conjunto e de transformação, com a adaptação e formatação para que o algoritmo seja aplicado com sucesso.

A mineração de dados com uso do Apriori foi realizada em cima do arquivo CSV gerado da consulta de escaneamento completo da tabela de junção do fato com as demais dimensões analisadas, após terei sido feitas as limpezas e transformações neste. Tal procedimento foi aplicado na IDE R Studio em um ambiente local.

3.1.4 Análise de dados

Utilizamos diversas operações matemática para podermos exprimir os melhores resultados, uma dessas operações foi a operação de soma para se obter os totais ou subtotais dos itens selecionados.

Figura 25 - Tabela de quantidade de inscritos segmentados pelo item cor e raça

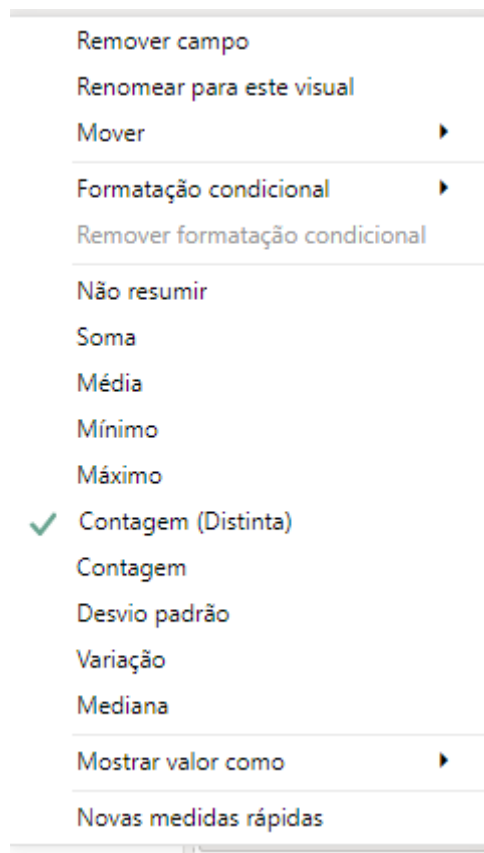
COR_RACA	Contagem de NU_INSCRICAO
Amarela	116162
Branca	1831779
Indígena	31756
Não declarado	103201
Parda	2364063
Preta	648309
Total	5095270

Fonte: Elaborado pelo autor (2021)

A segmentação dos dados se dar pela seleção das colunas que serão utilizadas, ao selecionar uma coluna, o Power BI realizar um agrupamento dados que estão repetidos na mesma coluna.

Na figura 25 há uma coluna que se inicia com o nome de contagem, isso ocorre por que a coluna de código de inscritos foi designada para ser a base de quantidade dos valores agregados da primeira coluna, a configuração de quantidade está na figura 26.

Figura 26 - Configuração de dados de uma coluna



Fonte: Elaborado pelo autor (2021)

Dentro destas opções pode se eleger a soma, média, mediana, entre outros.

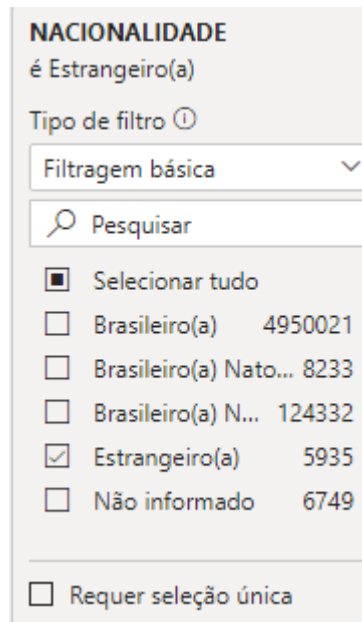
Dentro das configurações do Power BI pode se gerar uma visualização com filtro aplicado, para que se tenha um gráfico mais específico. A filtragem realizada na Figura 27 utilizou uma coluna já acionada ao gráfico da figura 27, neste caso só para visualizar o tipo de nacionalidade estrangeiro e não os demais.

Figura 27 - Tabela utilizando filtragem de dados

NACIONALIDADE	ST_CONCLUSAO	Contagem de NU_INSCRICAO
Estrangeiro(a)	Estou cursando e concluirei o Ensino Médio após 2019	618
Estrangeiro(a)	Estou cursando e concluirei o Ensino Médio em 2019	1780
Estrangeiro(a)	Já concluí o Ensino Médio	3511
Estrangeiro(a)	Não concluí e não estou cursando o Ensino Médio	26
Total		5935

Fonte: Elaborado pelo autor (2021)

Figura 28 - Filtragem de dados de uma coluna no Power BI



Fonte: Elaborado pelo autor (2021)

O cálculo de média aritmética foi utilizado para que fosse encontrado tendências e auxiliar nas informações extraídas da mineração.

A média aritmética é considerada uma medida de tendência central. Ela é obtida dividindo a soma dos números dados, pelos números somados (NOÉ, 2016).

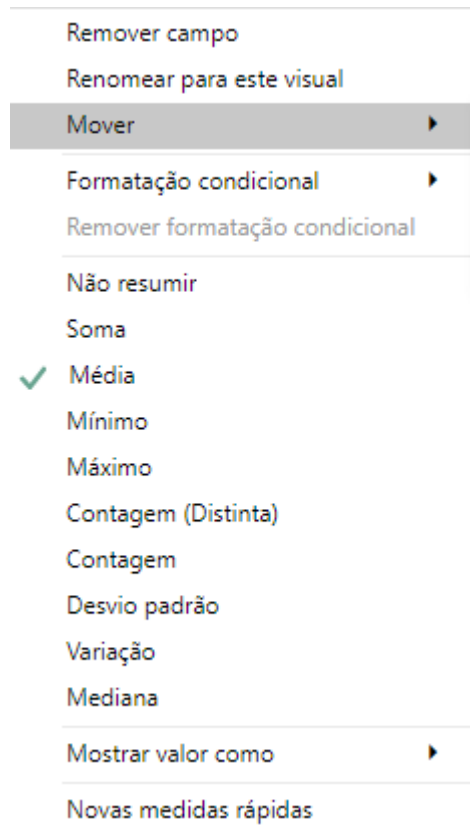
Para realizar esse procedimento dentro do Power BI tivemos de utilizar uma coluna que dados numéricos, depois selecionar a opção de média como está na figura 30 e o resultado está na figura 29.

Figura 29 - Tabela de média das notas da redação por faixa de idade

FX_IDADE	Média de NU_NOTA_REDACAO
Menor de 18 anos	595,66
Entre 18 a 24 anos	577,10
Entre 25 a 31 anos	533,97
Entre 31 a 36 anos	513,06
Entre 37 e 42 anos	487,59
Maior de 42 anos	461,65

Fonte: Elaborado pelo autor (2021)

Figura 30 - Configuração do Power BI para a definição da operação de média



Fonte: Elaborado pelo autor (2021)

A depender da ocasião elegimos utilizar a visualização dos gráficos em porcentagem para se obter uma ideia da proporção dos dados contidos no contexto daquela população eleita.

Na figura 31 pode ser visto como a porcentagem pode auxiliar na visualização das informações

Figura 31 - Tabela de porcentagem das faixas das notas da prova objetiva aplicada a gênero feminino

TP_SEXO	FX_MED_NOTA_PROVA_OBJ	%GT Contagem de NU_INSCRICAO
F		22,79%
F	200 a 400 pontos	2,32%
F	400 a 600 pontos	67,91%
F	Até 200 pontos	0,05%
F	Maior de 600 pontos	6,92%
Total		100,00%

Fonte: Elaborado pelo autor (2021)

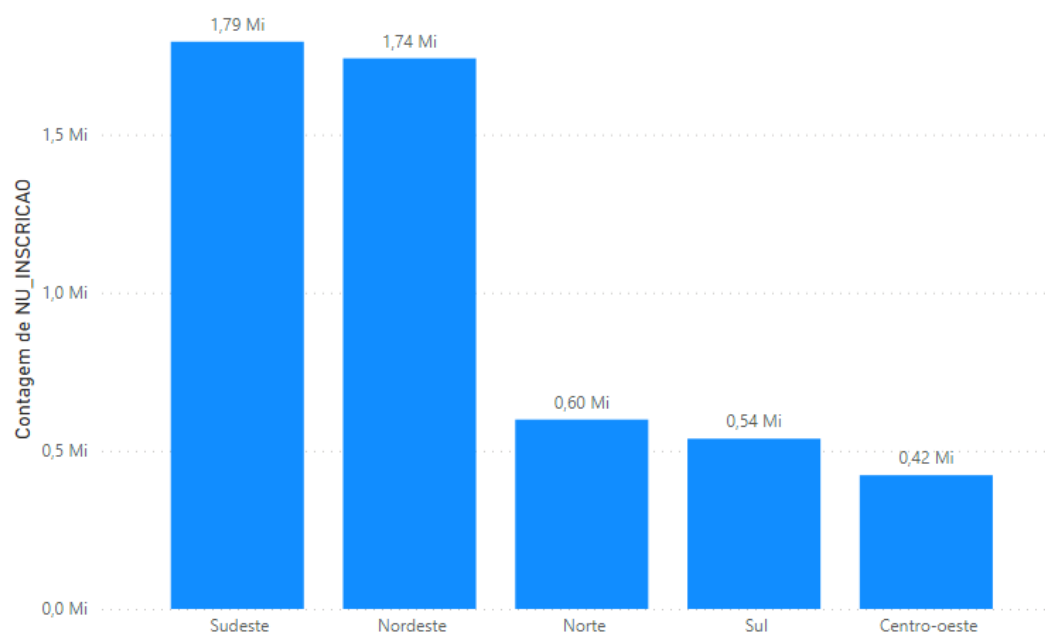
3.1.5 Visualização de dados

Os elementos visuais utilizados foram eleitos para facilitar o entendimento de todos, principalmente para os que são da área da educação, então tivemos o cuidado de ser o mais claro e mais conciso possível.

Utilizamos gráficos de colunas, mapa de árvore, cascata, funil, pizza, rosca, tabelas, mapas, cartões, segmentações, árvores hierárquicas.

Com os gráficos de colunas realizamos comparações entre os totais de conjunto de dados de uma coluna, como está na figura 32.

Figura 32 - Gráfico de colunas que mostra a comparação de totais de inscritos para cada região do Brasil



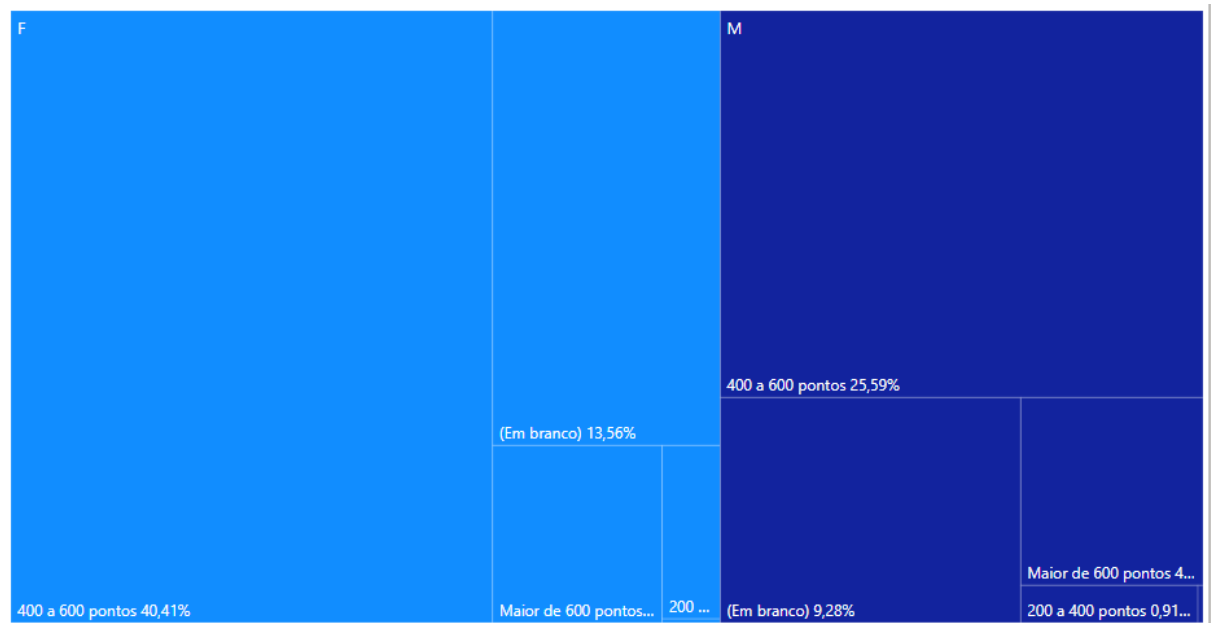
Fonte: Elaborador pelo autor (2021)

O *treemap* ou mapa de árvore foi utilizado para demonstrar a proporção de uma população de um conjunto de dados.

O gráfico do mapa fornece uma visão hierárquica dos dados, facilitando a identificação de padrões. Os ramos são representados por retângulos e cada sub-ramo é exibido como um retângulo menor. Os diagramas em árvore exibem categorias em cores e proximidade e podem exibir facilmente grandes quantidades de dados, enquanto outros tipos de diagramas são difíceis (Microsoft, 2021).

Na figura 33 é visto que a distribuição de informações é exibida segundo suas proporções, cada retângulo exprimi de forma visual a comparação das informações.

Figura 33 - Mapa de árvore exibindo uma comparação entre a quantidade de inscritos de distintos sexos e suas respectivas faixas notas

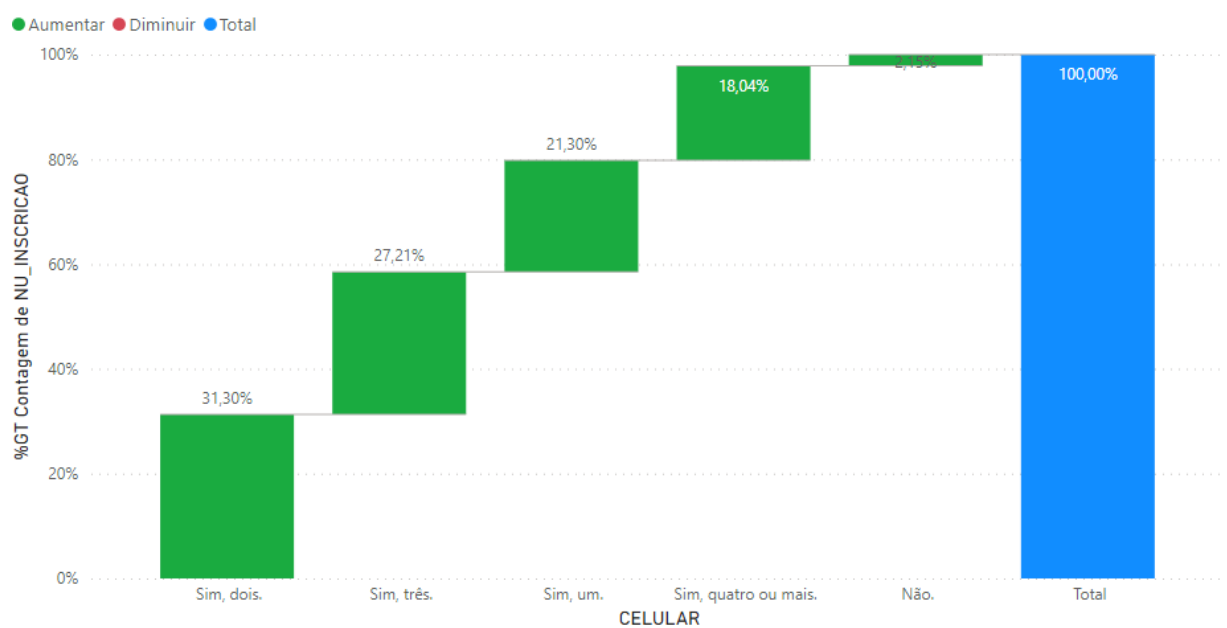


Fonte: Elaborado pelo autor (2021)

O gráfico em cascata mostra o total cumulativo ao adicionar ou subtrair valores. É útil para você entender como uma série de valores positivos e negativos afetam o valor inicial (Microsoft, 2021).

Na figura 34 vemos essa distribuição, e o valor final sendo montado ou justificado.

Figura 34 - Gráfico de cascata que exibi as distribuições da proporção das quantidades de celulares que há na residência dos inscritos



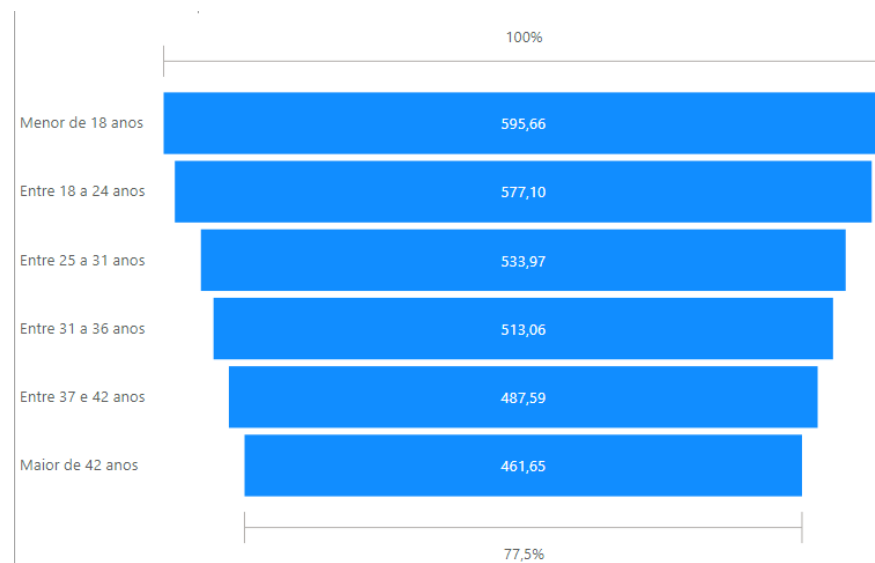
Fonte: Elaborado pelo autor (2021)

O gráfico de funil, no auxiliar no entender os valores de etapas dentro de um encadeamento.

O gráfico de funil mostra o valor de cada etapa do processo. Normalmente, esses valores diminuem gradualmente, fazendo com que as barras pareçam funis. (MICROSOFT, 2021)

Na figura 35, pode ver como está distribuído o gráfico.

Figura 35 - Visualização de funil das informações das faixas de idade com a nota média da redação.



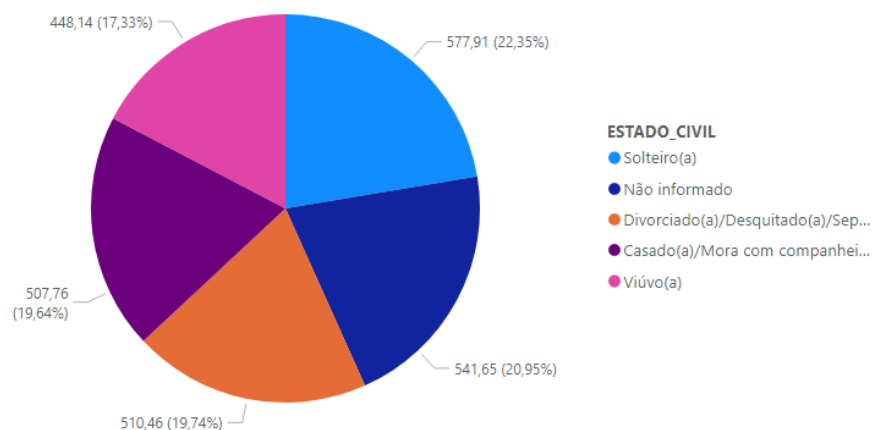
Fonte: Elaborado pelo autor (2021)

O gráfico de pizza foi utilizado devido sua facilidade de ser interpretado.

O gráfico de pizza é usado para exibir a proporção dos dados da categoria, e o tamanho de cada parte representa a proporção de cada categoria (ARCGIS, 2021).

Na figura 36 podemos ver um exemplo do gráfico de pizza aplicado.

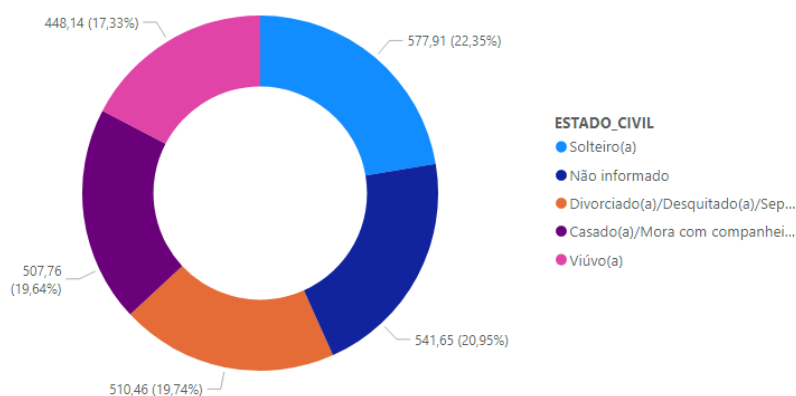
Figura 36 -Gráfico de pizza exibindo a média da nota da redação segmentado pelo estado civil dos participantes do exame



Fonte: Elaborado pelo autor (2021)

No gráfico de rocas vimos algo similar ao de pizza, como está na figura 37, ele também ótimo para abstração do entendimento da informação.

Figura 37 -Gráfico de rosca exibindo a média da nota da redação segmentado pelo estado civil dos participantes do exame



Fonte: Elaborado pelo autor (2021)

A visualização em tabelas do Power BI é uma forma simples de visualizar os dados, geralmente de forma agrupada, como na figura 38.

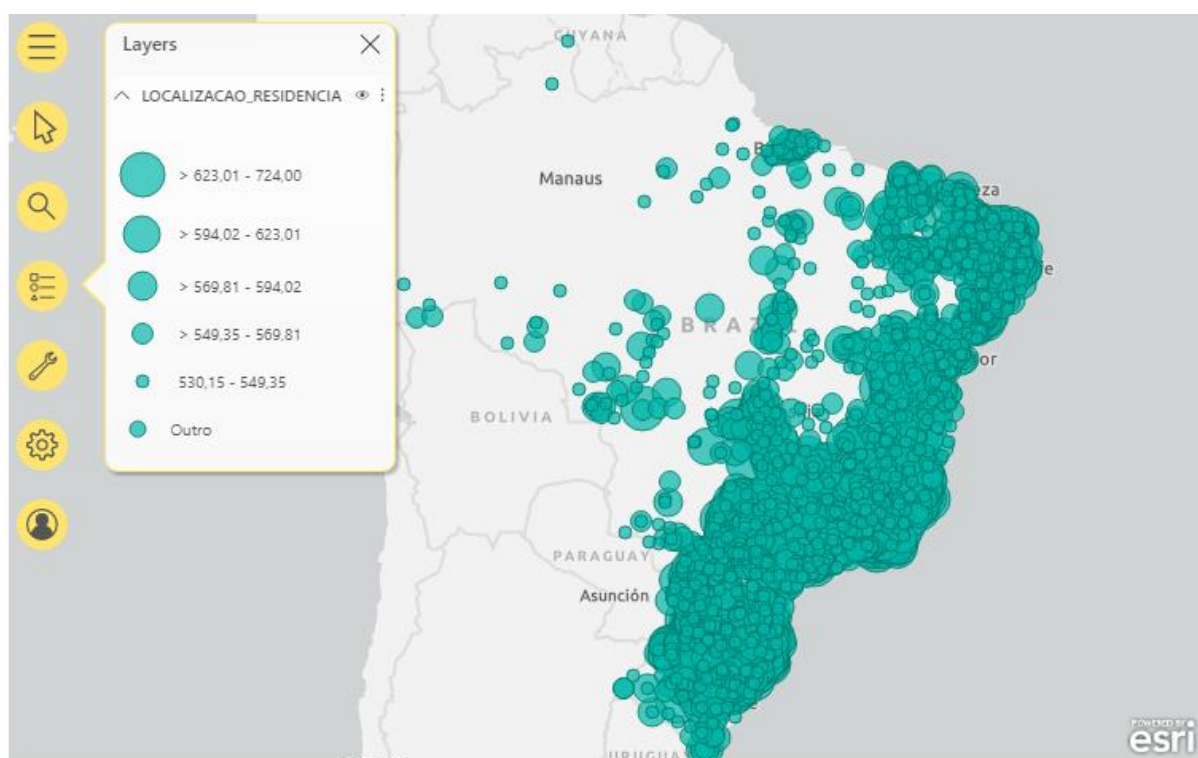
Figura 38 - Tabela comparativa sobre DVD e TV por assinatura junto da média da nota da redação

TV_ASSINATURA	DVD	Média de NU_NOTA_REDACAO
Sim.	Sim.	645,75
Sim.	Não.	612,84
Não.	Sim.	561,55
Não.	Não.	551,10
Total		571,19

Fonte: Elaborado pelo autor (2021)

A inserção da visualização de mapa criar uma facilidade para encontrar informações e associá-las a determinada região, podendo ser visto de uma forma mais abrangente ou mais específico através do zoom aplicado a imagem. Na figura 39 pode ser visto essa distribuição, que cada representa um agrupamento de notas criado pelo próprio Power BI.

Figura 39 - média de notas da redação distribuídas por cidades do Brasil.



Fonte: Elaborado pelo Autor (2021)

Os cartões são formas de visualizar uma informação em destaque, ele só exibe um dado, ao contrário dos outros gráficos que procuram associar e comparar valores, o cartão tem a função de destacar um valor, seja o resulta de uma soma, média ou até o maior valor de um

determinado conjunto de número. Nós utilizamos o gráfico para exibir a quantidade de inscritos no exame, como está na figura 40

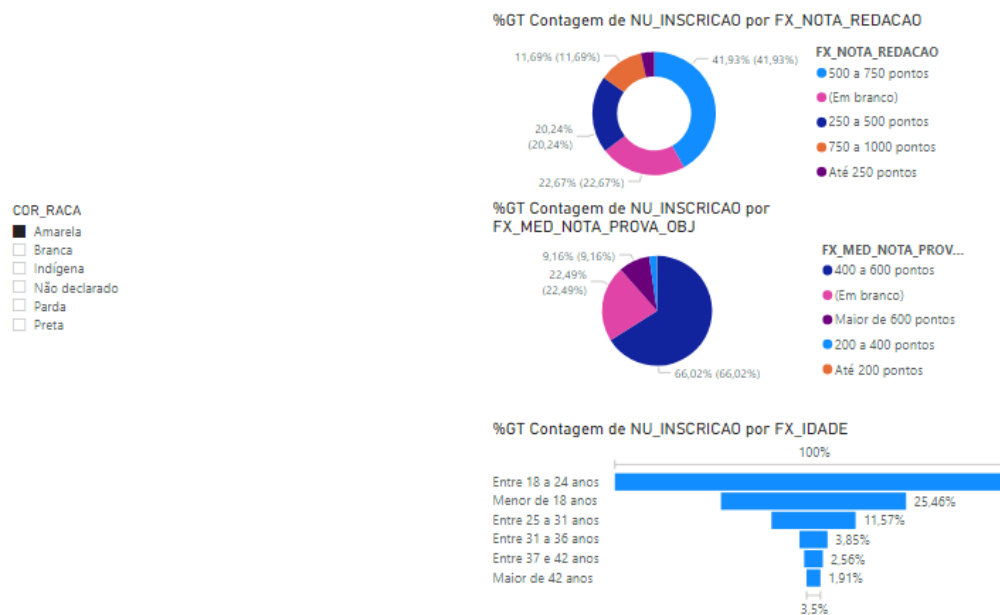
Figura 40 - Visualização de cartão, destacando o número de inscritos no exame



Fonte: Elaborado pelo próprio autor

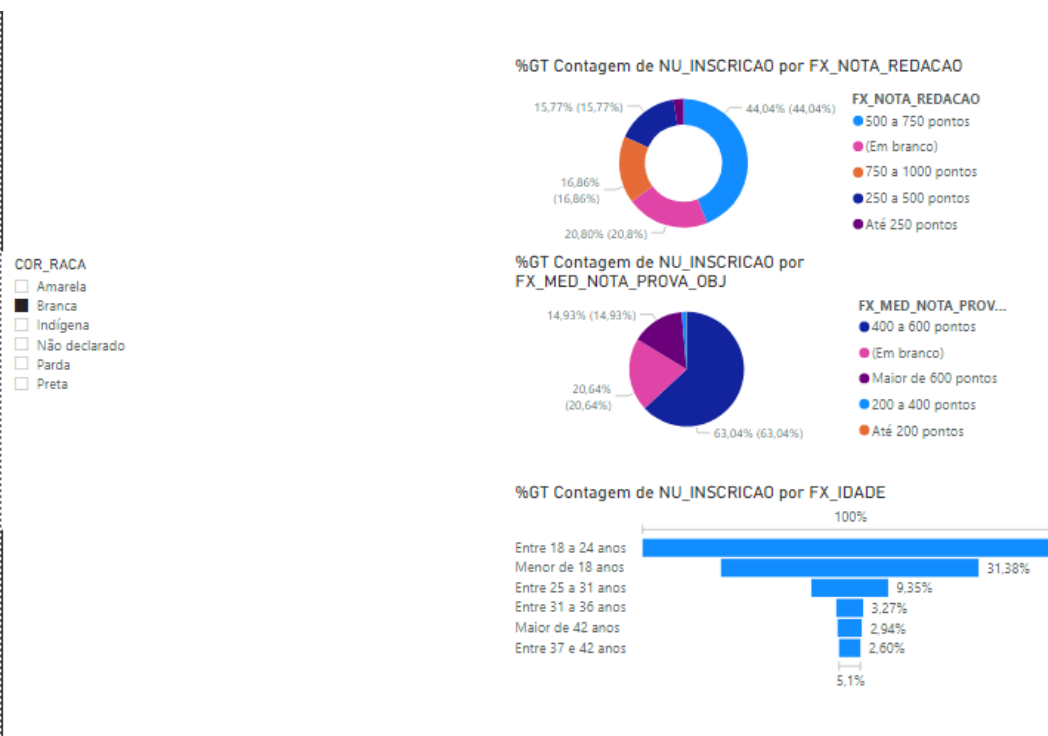
A segmentação é um dos visualizadores que mais proporciona a interação do usuário com relatório, pois a segmentação sugestiona ao usuário experimentar uma mudança de visualização dos dados gerais e aplicá-lo para determinado grupo de dados e após isso ele poderá verificar outro grupo de dados, que outrora os visuais estavam aplicados apenas para dados em geral, entretendo agora, está aplicado a dados mais específicos. O Power BI possui um editor interações, no qual permite os gráficos e visuais interagirem conforme o clique do usuário nos visuais, caso os dados contenham uma correlação entre si, o Power BI destacá-los e na segmentação não é diferente já que o intuito dessa ferramenta é justamente esse. Na figura 41 e 42 vemos a diferença entre as interações

Figura 41 - Segmentação dos gráficos que contêm as informações de nota da redação



Fonte: Elaborado pelo autor (2021)

Figura 42 - Segmentação com a opção trocada e a interação dos visuais



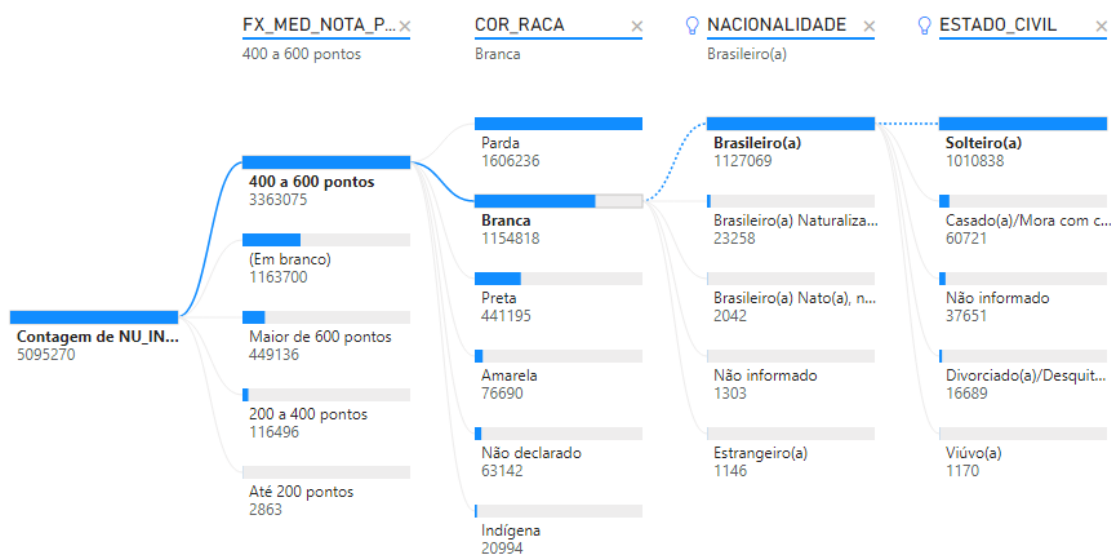
Fonte: Elaborado pelo autor (2021)

O gráfico de árvore hierárquica nos auxilia na verificação dos resultados da mineração, nele podemos fazer a verificação por quantidade de inscritos, ou pela qualidade da média das notas.

A aparência da árvore de decomposição no Power BI permite que os dados sejam exibidos em várias dimensões. Ele agrega automaticamente os dados e permite pesquisas detalhadas de suas dimensões em qualquer ordem. É também a visualização da IA (inteligência artificial), de modo que a próxima dimensão pode ser encontrada para pesquisa detalhada com base em condições específicas. Isso o torna uma ferramenta valiosa para exploração ad hoc e análise de causa raiz (MIHART; OLPROD, 2020).

Na figura 43 podemos ver cada caminho percorrido pela árvore, baseando-se na média da nota, com essa ferramenta somos capazes de exibir os valores que a mineração expõe através da frequência, mas também podemos realizar comparações utilizando as médias para termos não só a quantidade, mas também a qualidade.

Figura 43 - Gráfico de árvore hierárquica com os dados dos participantes solteiros, brancos, brasileiros com faixa de nota entre 400 a 600 pontos na prova objetiva, mensurando pela quantidade de inscrito por atributos



Fonte: Elaborado pelo autor (2021)

3.2 ENTENDIMENTO DO NEGÓCIO

O entendimento do domínio de negócio da educação de base, mais especificamente o ensino médio, foi pautado na entrevista de especialistas em pedagogia e ensino com a utilização de questionários on-line. Primeiramente, perguntou-se sobre uso de dados educacionais e

pedagógicos para avaliação de desempenho, posteriormente foi perguntado especificamente a professores de ensino médio sobre associações e correlações interessantes nos dados expostos a eles, que compõem a base de micro dados do Enem.

3.2.1 Questionários para entendimento do uso de dados pedagógicos para avaliação de ambientes educacionais

Neste questionário, foram feitas perguntas com foco em extrair dos profissionais quais dados são mais apropriados para determinados propósitos avaliativos. Segmentou-se as perguntas através de níveis hierárquicos de atuação, divididos entre professores, coordenadores, diretores e pedagogos em geral. O foco foi perguntar e tentar obter informações que fugissem ao mero caráter quantitativo e fossem mais diversificadas, como dados pessoais ou sociais dos alunos. No que tange aos dados quantitativos em si, perguntou-se como analisá-los em agrupamentos.

Como o foco do trabalho precisou ser mudado, em virtude da impossibilidade de se conseguir bases de dados de escolas para analisar, então o feedback de especialistas neste cenário serviu como base para definirmos quais as variáveis eles acham mais relevantes para se atentar quando se realiza avaliação de fatores por trás do desempenho de alunos e como podem apoiar tomadas de decisão orientada a estes dados.

3.2.2 Questionário para definição e avaliação de conjuntos de itens pertinentes para associação no contexto do desempenho e dos perfis socioeconômicos dos participantes do Enem 2019

No segundo questionário aplicado, o foco delimitou-se em professores de ensino médio, preparadores de cursos pré-vestibular e profissionais de assistência social para levantar sua percepção especializada acerca de fatores diversos correspondentes aos dados presentes na base de dados no sentido de descobrir quais deles podem ser associados ao desempenho de um participante do Enem.

A ideia é apoiar a análise dos resultados na etapa de mineração de dados, para identificar quais as variáveis referentes a conjuntos de dados são mais relevantes para analisar sob a ótica dos profissionais da área e nas fases da mineração em diferentes limiares de probabilidade, comparar a visão especialista aos resultados obtidos para consultá-los na interpretação deles.

3.2.3 Feedback para interpretação e avaliação dos resultados

Neste último questionário, foi exposto aos especialistas uma planilha contendo o resultado geral das associações mais comuns em todos os testes, com os parâmetros de suporte, confiança e alavancagem para cada um deles. Foi solicitado a interpretação de cada regra encontrada e o porquê dos seus limiares probabilísticos.

3.3 KDD

Neste capítulo, são descritas as etapas do KDD e o que foi desenvolvido em cada uma delas. Esta etapa do trabalho é onde foi aplicado o algoritmo de mineração de dados para análise avaliativa diagnóstica do ENEM 2019.

3.3.1 Coleta de dados

Os dados foram coletados do portal do Instituto de Pesquisas Econômicas Anísio Teixeira, o INEP, através do download de um arquivo ZIP em seu website. O ZIP continha a base bruta de micro dados em formato CSV, as informações de itens de prova no formato XLSX e o dicionário da base de micro dados em formato XLSX e ODS.

A base bruta de dados e o dicionário foram carregados para um repositório de armazenamento em nuvem da Azure em arquitetura data lake, que é o Data Lake Storage Gen 1, através da ferramenta de gerenciamento Azure Storage Explorer, devido a limitação da interface web no endereço SaaS do ADLS de até 2GB por envio. O arquivo fonte bruto foi inserido na pasta RawData, para consumação pelos pipelines de ingestão do ADF.

Os arquivos de cada tabela do DM foram carregados para uma pasta TransformedModelRDW para serem carregados para o cubo OLAP do AAS. Já o arquivo para mineração foi carregado em uma pasta chamada StaggeringPreProcessed, após as transformações.

3.3.2 Pré-processamento

A etapa de pré-processamento é caracterizada pela seleção dos dados do modelo, a limpeza dos dados sem utilidade para o propósito da mineração pela remoção de colunas utilizadas no OLAP, porém desnecessária no algoritmo e o tratamento de redundância pela modificação dos dados para adição de referência.

3.3.2.1 Seleção dos dados

Foram selecionados todos os dados da base, com exceção daqueles que compõe as respostas e o gabarito da prova, bem como o ano de conclusão do ensino médio e o ano referente a prova.

3.3.2.2 Limpeza de dados

No modelo de dados para mineração foram retiradas as colunas de códigos de regionalidade, que caracteriza uma redundância desnecessária por existir os dados de localidade especificamente e código da escola também fora retirado, pois avalia-se aspectos gerais do aluno sem delimitar informações pessoais específicas. Através da análise dos tipos de ensino e de escola, é possível uma avaliação satisfatória.

3.3.2.3 Eliminação de redundância dos dados de valores idênticos: Inserção de identificador de células

De acordo com o dicionário de dados transformados, foi realizada a atualização dos dados para criação do modelo pré-processado. De acordo com as transcrições do dicionário original para o novo, foi realizado o update de cada coluna com tratamento de exceção envolvendo submissão e possibilidade de desfazer a transação a efetuar para persistência dos dados.

Figura 44 - Pré-processamento – Tratamento de redundâncias - Estado Civil

```
--Marital Status
-- Not Informed | Single | Married | Divorced or Separated | Widowed
update jft.Join_FactDim_InfoProva_InfoPessoa_STEM_QSE
set TP_ESTADO_CIVIL = 'MS - NIF'
where TP_ESTADO_CIVIL = '0'

update jft.Join_FactDim_InfoProva_InfoPessoa_STEM_QSE
set TP_ESTADO_CIVIL = 'MS - SIN'
where TP_ESTADO_CIVIL = '1'

update jft.Join_FactDim_InfoProva_InfoPessoa_STEM_QSE
set TP_ESTADO_CIVIL = 'MS - MAR'
where TP_ESTADO_CIVIL = '2'

update jft.Join_FactDim_InfoProva_InfoPessoa_STEM_QSE
set TP_ESTADO_CIVIL = 'MS - DIV'
where TP_ESTADO_CIVIL = '3'

update jft.Join_FactDim_InfoProva_InfoPessoa_STEM_QSE
set TP_ESTADO_CIVIL = 'MS - WID'
where TP_ESTADO_CIVIL = '4'
```

Fonte: Elaborado pelo autor (2021)

Figura 45- Pré-processamento – Tratamento de redundâncias - Cor ou raça

```
-- RaceColour
-- NonDeclared | White | Black | Brown | Yellow | Indigenous
update jft.Join_FactDim_InfoProva_InfoPessoa_STEM_QSE
set TP_COR_RACA = 'RC - NDC'
where TP_COR_RACA = '0'

update jft.Join_FactDim_InfoProva_InfoPessoa_STEM_QSE
set TP_COR_RACA = 'RC - WHI'
where TP_COR_RACA = '1'

update jft.Join_FactDim_InfoProva_InfoPessoa_STEM_QSE
set TP_COR_RACA = 'RC - BLK'
where TP_COR_RACA = '2'

update jft.Join_FactDim_InfoProva_InfoPessoa_STEM_QSE
set TP_COR_RACA = 'RC - BRW'
where TP_COR_RACA = '3'

update jft.Join_FactDim_InfoProva_InfoPessoa_STEM_QSE
set TP_COR_RACA = 'RC - YLW'
where TP_COR_RACA = '4'

update jft.Join_FactDim_InfoProva_InfoPessoa_STEM_QSE
set TP_COR_RACA = 'RC - IDN'
where TP_COR_RACA = '5'
```

Fonte: Elaborado pelo autor (2021)

Figura 46 - Pré-processamento – Tratamento de redundâncias - Questionários Socioeconômicos

```
-- Q001 - FATHER'S SCHOOL LEVEL
update jft.Join_FactDim_InfoProva_InfoPessoa_STEM_QSE
set Q001 = 'SEF_FSL - NS'
where Q001 = 'A'

update jft.Join_FactDim_InfoProva_InfoPessoa_STEM_QSE
set Q001 = 'SEF_FSL - NCEFS'
where Q001 = 'B'

update jft.Join_FactDim_InfoProva_InfoPessoa_STEM_QSE
set Q001 = 'SEF_FSL - NCESS'
where Q001 = 'C'

update jft.Join_FactDim_InfoProva_InfoPessoa_STEM_QSE
set Q001 = 'SEF_FSL - NCHS'
where Q001 = 'D'

update jft.Join_FactDim_InfoProva_InfoPessoa_STEM_QSE
set Q001 = 'SEF_FSL - CHSNC'
where Q001 = 'E'

update jft.Join_FactDim_InfoProva_InfoPessoa_STEM_QSE
set Q001 = 'SEF_FSL - CCNP'
where Q001 = 'F'

update jft.Join_FactDim_InfoProva_InfoPessoa_STEM_QSE
set Q001 = 'SEF_FSL - CPGS'
where Q001 = 'G'
```

Fonte: Elaborado pelo autor (2021)

3.3.3 Transformação

As transformações foram realizadas com o emprego da linguagem T-SQL para atualizar os campos mediante as condições necessárias especificadas no dicionário da transformação e pré-processamento. Em suma, foca na criação da média das notas a partir do cálculo das notas de todas as áreas de conhecimento(knowledge areas), a transformação da nota da redação em campo textual(varchar) com a conversão de tipo para comparação de faixa numérica(cast decimal). Para retirada dos nulos, utilizou-se o validador ISNULL. O modelo do conjunto de dados resultante possui todos os dados padronizados conforme pré-estabelecido no dicionário.

3.3.3.1 Cálculo de média das notas

As médias das notas foram calculadas através de um UPDATE que realiza Join implícito do atributo das notas

Figura 47 - Transformação - Cálculo de média das notas

```
-- NU_MEDIA_TOTAL_AREAS

-- Cast and calculate AVG grades
update jft.Join_FactDim_InfoProva_InfoPessoa_STEM_QSE
set NU_MEDIA_TOTAL_AREAS = totalgrades from [jft].[Join_FactDim_InfoProva_InfoPessoa_STEM_QSE] join(select
AVG((cast(NU_NOTA_CN as decimal(5,2))+
cast(NU_NOTA_CH as decimal(5,2))+
cast(NU_NOTA_LC as decimal(5,2))+
cast(NU_NOTA_MT as decimal(5,2))
)/5) as totalgrades, NU_INSCRICAO_FACT_DIMS
from [jft].[Join_FactDim_InfoProva_InfoPessoa_STEM_QSE]
group by NU_INSCRICAO_FACT_DIMS) agregation
on [jft].[Join_FactDim_InfoProva_InfoPessoa_STEM_QSE].[NU_INSCRICAO_FACT_DIMS] = agregation.NU_INSCRICAO_FACT_DIMS

update jft.Join_FactDim_InfoProva_InfoPessoa_STEM_QSE
set NU_MEDIA_TOTAL_AREAS = 'AVG_KA_NP_NULL'
where NU_MEDIA_TOTAL_AREAS is null
```

Fonte: Elaborado pelo autor (2021)

3.3.3.2 Remoção de nulos

Os nulos foram removidos com transformações via atualização, com a utilização do comparador ISNULL na clausula WHERE.

Figura 48 - Transformação - Tratamento de nulos – Idade - 2

```
-- Exams nulls CN - CH - MT - LP
update jft.Join_FactDim_InfoProva_InfoPessoa_STEM_QSE
set CO_PROVA_CN = 'NSEP - NP_NULL'
where CO_PROVA_CN is null

update jft.Join_FactDim_InfoProva_InfoPessoa_STEM_QSE
set CO_PROVA_CH = 'HSEP - NP_NULL'
where CO_PROVA_CH is null

update jft.Join_FactDim_InfoProva_InfoPessoa_STEM_QSE
set CO_PROVA_LC = 'LCEP - NP_NULL'
where CO_PROVA_LC is null

update jft.Join_FactDim_InfoProva_InfoPessoa_STEM_QSE
set CO_PROVA_MT = 'MCEP - NP_NULL'
where CO_PROVA_MT is null

update jft.Join_FactDim_InfoProva_InfoPessoa_STEM_QSE
set TP_STATUS_REDACAO = 'MCEP - NP_NULL'
where TP_STATUS_REDACAO is null
```

Fonte: Elaborado pelo autor (2021)

3.3.3.3 Discretização de valores contínuos

Foram efetuados updates com sub consultas com a causa de comparação entre valores numéricos between para obtenção das faixas categóricas discretizadas das médias das notas das áreas de conhecimento, das competências e da nota geral da redação.

Como o formato das notas está em varchar, isto é, um campo textual, aplicou-se a função T-SQL de conversão de tipos Cast para realizar a tradução em tempo de execução no processo de consulta. Para armazenar a média de notas de todas as áreas, foi criada a coluna em formato decimal(real) denominada NU_MEDIA_TOTAL_AREAS para capturar o resultado do cálculo. Todas as inserções foram carregadas por meio de sub consulta em cada tabela de referência para efetivar as alterações, com aplicação do conversor de tipo na análise exploratória dos dados de cada coluna para viabilizar o uso do operador between.

3.3.3.4 Agrupamento classificatório de valores discretos

Foi realizada a sumarização categórica dos valores de idade na coluna NU_IDADE, sendo dividido entre grupos descritos no dicionário de dados transformados, que compreendem em idades menores que 18, maiores que 42 e, entre eles, variações de 6 em 6 anos.

Figura 49 - base de dados transformada - Classificação de discretos – Idade - 1

```
update jft.Join_FactDim_InfoProva_InfoPessoa_STEM_QSE
set NU_IDADE = 'AGE_R LT18'
where CAST(NU_IDADE AS tinyint)<18
AND NU_IDADE not in('AGE_R LT18','AGE_R GT42','AGE_R [18 - 24]','AGE_R [25 - 30]','AGE_R [31 - 36]','AGE_R [37 - 42]')

update jft.Join_FactDim_InfoProva_InfoPessoa_STEM_QSE
set NU_IDADE = 'AGE_R [18 - 24]'
where CAST(NU_IDADE AS tinyint) BETWEEN 17 AND 25
AND NU_IDADE not in('AGE_R LT18','AGE_R GT42','AGE_R [18 - 24]','AGE_R [25 - 30]','AGE_R [31 - 36]','AGE_R [37 - 42]')
```

Fonte: Elaborado pelo autor (2021)

Figura 50 - base de dados transformada - Classificação de discretos – Idade - 2

```
update jft.Join_FactDim_InfoProva_InfoPessoa_STEM_QSE
set NU_IDADE = 'AGE_R [37 - 42]'
where CAST(NU_IDADE AS tinyint) BETWEEN 36 AND 43
AND NU_IDADE not in('AGE_R LT18','AGE_R GT42','AGE_R [18 - 24]','AGE_R [25 - 30]','AGE_R [31 - 36]','AGE_R [37 - 42]')

update jft.Join_FactDim_InfoProva_InfoPessoa_STEM_QSE
set NU_IDADE = 'AGE_R GT42'
where CAST(NU_IDADE AS tinyint) > 42
AND NU_IDADE not in('AGE_R LT18','AGE_R GT42','AGE_R [18 - 24]','AGE_R [25 - 30]','AGE_R [31 - 36]','AGE_R [37 - 42]')
```

Fonte: Elaborado pelo autor (2021)

3.3.3.5 Transformação transacional dos itens

Foi realizada a transformação dos conjuntos de dados em um formato transacional do tipo "cesta" (basket), que considera cada linha como um registro de transação único com uma identificação e cada coluna como um item. No momento da leitura das transações do arquivo transformado, foi retirado a coluna de cabeçalho, para a não poluição dos conjuntos de itens durante a aplicação do algoritmo Apriori.

Figura 51 - leitura de transações e remoção de cabeçalhos do arquivo contendo questionários de situação familiar

```
SEEP_FS=read.transactions("~path/TCC/Bases/Transformed/ExamFactSEF_FamilySituation.csv",format='basket',sep=",")
SEEP_FS<-SEEP_FS[-1,]
```

Fonte: Elaborado pelo autor (2021)

3.3.4 Mineração de dados

Nesta etapa, é descrito o processo de aplicação do algoritmo apriori e a estratégia de segmentação dos grupos de dados a ser minerados.

3.3.4.1 Descrição do dicionário de dados pré-processados

O dicionário para transformação foi transposto para a língua inglesa e transcreve o nome das colunas para esta, adicionando a informação específica e a quantidade, se houver, no dado em si. Segue-se padrão de acordo com o item definido. A primeira parte de cada item descreve, de 2 a quatro siglas, o campo tratado em inglês e o restante descreve o seu conteúdo. Como exemplos, o dado MS – SIN significa *marital status - single*, ou estado civil - solteiro. SEF – RHB – Y1BR significa *social economic form – residence has bathroom – yes 1 bathroom*, ou questionário socioeconômico - a residência tem banheiro(s) - sim 1 banheiro. Est padrão é seguido para todas as transformações e o pré-processamento de redundâncias, a variar da especificidade dos dados, se são quantitativos ou apenas descrições qualitativas. Para os dados nulos, é transcrito o motivo da nulificação, como por exemplo OS – NA_ACH, que significa *operational situation(school) – not applied_already concluded highschool*, ou situação de operação (da escola) - não aplicável já concluiu o ensino médio.

3.3.5 Modelo de dados para mineração

O modelo gerado para mineração contém apenas a nota da redação, proveniente da tabela fato, o número de inscrição da tabela de Join do fato com as dimensões, a média das notas que estavam na “fato” e as informações das tabelas de dimensão, exceto de inclusão, que não é o foco analítico do trabalho. Na figura consta o modelo gerado.

Figura 52 - base de dados transformada - 2

Join FactDim InfoProva InfoPessoa STEM QSE (IT)
NU_INSCRICAO_FACT_DIMS
NU_NOTA_REDACAO
NO_MUNICIPIO_PROVA
SG_UF_PROVA
TP_PRESENCA_CN
TP_PRESENCA_CH
TP_PRESENCA_LC
TP_PRESENCA_MT
CO_PROVA_CN
CO_PROVA_CH
CO_PROVA_LC
CO_PROVA_MT
TP_LINGUA
TP_STATUS_REDACAO
NO_MUNICIPIO_RESIDENCIA
SG_UF_RESIDENCIA
NU_IDADE
TP_SEXO
TP_ESTADO_CIVIL
TP_COR_RACA
TP_NACIONALIDADE
NO_MUNICIPIO_NASCIMENTO
SG_UF_NASCIMENTO
TP_ST_CONCLUSAO
TP_ESCOLA
TP_ENSINO
IN_TREINERO
NO_MUNICIPIO_ESC
SG_UF_ESC
TP_DEPENDENCIA_ADM_ESC
TP_LOCALIZACAO_ESC
TP_ST_FUNC_ESC
Q001
Q002
Q003
Q004
Q005
Q006
Q007
Q008
Q009
Q010
Q011
Q012
Q013
Q014
Q015
Q016
Q017
Q018
Q019
Q020
Q021
Q022
Q023
Q024
Q025
NU_MEDIA_TOTAL_AREAS

Fonte: Elaborado pelo autor (2021)

Figura 53 - base de dados transformada - 1

1	SELECT	[NU_INSCRICAO_FACT_DIMS]
2		, [NU_NOTA_REDACAO]
3		, [NO_MUNICIPIO_PROVA]
4		, [SG_UF_PROVA]
5		, [TP_PRESENCA_CN]
6		, [TP_PRESENCA_CH]
7		, [TP_PRESENCA_LC]
8		, [TP_PRESENCA_MT]
9		, [CO_PROVA_CN]
10		, [CO_PROVA_CH]
11		, [CO_PROVA_LC]

	NU_INSCRICAO_FACT_DIMS	NU_NOTA_REDACAO	NO_MUNICIPIO_PROVA	SG_UF_PROVA	TP_PRESENCA_CN	TP_PRESENCA_CH	TP_PRESENCA_LC	TP_PRESENCA_MT	CO_PROVA_CN	CO_PROVA_CH	CO_PRO
1	190001004627	ETGI -]750 - 1000]	ECN - Santarém	EFU - PA	NSP - PST	HSP - PST	LCP - PST	MEP - PST	NSEC - NRM	HSEC - NRM	LCEC
2	190001004628	ETGI -]500 - 750]	ECN - Pôr do Rio	EFU - PA	NSP - PST	HSP - PST	LCP - PST	MEP - PST	NSEC - NRM	HSEC - NRM	LCEC
3	190001004629	ETGI -]500 - 750]	ECN - Paragominas	EFU - PA	NSP - PST	HSP - PST	LCP - PST	MEP - PST	NSEC - NRM	HSEC - NRM	LCEC
4	190001004630	ESSAY - NP_NULL	ECN - São Sebast.	EFU - PA	NSP - MIS	HSP - MIS	LCP - MIS	MEP - MIS	NSEP - NULL	HSEP - NULL	LCEC
5	190001004631	ESSAY - NP_NULL	ECN - Juruti	EFU - PA	NSP - MIS	HSP - MIS	LCP - MIS	MEP - MIS	NSEP - NULL	HSEP - NULL	LCEC
6	190001004632	ETGI -]500 - 750]	ECN - Belém	EFU - PA	NSP - PST	HSP - PST	LCP - PST	MEP - PST	NSEC - NRM	HSEC - NRM	LCEC
7	190001004633	ETGI -]250 - 500]	ECN - Marabá	EFU - PA	NSP - PST	HSP - PST	LCP - PST	MEP - PST	NSEC - NRM	HSEC - NRM	LCEC
8	190001004634	ETGI -]250 - 500]	ECN - Belém	EFU - PA	NSP - PST	HSP - PST	LCP - PST	MEP - PST	NSEC - NRM	HSEC - NRM	LCEC
9	190001004635	ETGI -]250 - 500]	ECN - Itaituba	EFU - PA	NSP - PST	HSP - PST	LCP - PST	MEP - PST	NSEC - NRM	HSEC - NRM	LCEC
10	190001004636	ETGI -]500 - 750]	ECN - São Miguel.	EFU - PA	NSP - PST	HSP - PST	LCP - PST	MEP - PST	NSEC - NRM	HSEC - NRM	LCEC
11	190001004637	ETGI -]750 - 1000]	ECN - São Miguel.	EFU - PA	NSP - PST	HSP - PST	LCP - PST	MEP - PST	NSEC - NRM	HSEC - NRM	LCEC
12	190001004638	ESSAY - NP_NULL	ECN - Tucuruí	EFU - PA	NSP - MIS	HSP - MIS	LCP - MIS	MEP - MIS	NSEP - NULL	HSEP - NULL	LCEC
13	190001004639	ESSAY - NP_NULL	ECN - Ananindeua	EFU - PA	NSP - MIS	HSP - MIS	LCP - MIS	MEP - MIS	NSEP - NULL	HSEP - NULL	LCEC

Fonte: Elaborado pelo autor (2021)

Figura 54 - base de dados transformada - 2

Run
Cancel
Disconnect
Change Connection
EnemInkedServerSQL
Explain
Enable SQLCMD
Export as Notebook

```

1 SELECT [NU_INSCRICAO_FAC_DIMS]
2 , [NU_NOTA_REDAcao]
3 , [NU_INSCRICAO_PROVA]
4 , [SE_UF_PROVA]
5 , [TP_PRESENCA_CH]
6 , [TP_PRESENCA_CH]
7 , [TP_PRESENCA_LC]
8 , [TP_PRESENCA_MT]
9 , [CO_PROVA_CH]
10 , [CO_PROVA_CH]
11 , [CO_PROVA_LC]

```

Results

Messages

Q016	Q017	Q018	Q019	Q020	Q021	Q022	Q023	Q024	Q025	NU_MEDIA_TOTAL_AREAS
SEF_RHQM_HRMQ	SEF_RHQM_HRMQ	SEF_RHCV - N	SEF_RHCT_Y1CT	SEF_RHDO - Y	SEF_RHST - N	SEF_RHCP_Y3CP	SEF_RHLP - N	SEF_RHC_Y1C	SEF_RHIA - N	AVG_KA - [200 - 400]
SEF_RHQM_HRMQ	SEF_RHQM_HRMQ	SEF_RHCV - N	SEF_RHCT_Y1CT	SEF_RHDO - Y	SEF_RHST - N	SEF_RHCP_Y3CP	SEF_RHLP - N	SEF_RHC_HMC	SEF_RHIA - N	AVG_KA - [200 - 400]
SEF_RHQM_HRMQ	SEF_RHQM_HRMQ	SEF_RHCV - N	SEF_RHCT_Y1CT	SEF_RHDO - Y	SEF_RHST - Y	SEF_RHCP_Y3CP	SEF_RHLP - N	SEF_RHC_Y2C	SEF_RHIA - Y	AVG_KA - [200 - 600]
SEF_RHQM_HRMQ	SEF_RHQM_HRMQ	SEF_RHCV - N	SEF_RHCT_Y1CT	SEF_RHDO - N	SEF_RHST - N	SEF_RHCP_Y3CP	SEF_RHLP - N	SEF_RHC_HMC	SEF_RHIA - N	AVG_KA_NP_NULL
SEF_RHQM_HRMQ	SEF_RHQM_HRMQ	SEF_RHCV - N	SEF_RHCT_Y1CT	SEF_RHDO - N	SEF_RHST - N	SEF_RHCP_Y3CP	SEF_RHLP - N	SEF_RHC_HMC	SEF_RHIA - N	AVG_KA_NP_NULL
SEF_RHQM_HRMQ	SEF_RHQM_HRMQ	SEF_RHCV - N	SEF_RHCT_Y1CT	SEF_RHDO - N	SEF_RHST - N	SEF_RHCP_Y3CP	SEF_RHLP - N	SEF_RHC_HMC	SEF_RHIA - N	AVG_KA - [400 - 600]
SEF_RHQM_HRMQ	SEF_RHQM_HRMQ	SEF_RHCV - N	SEF_RHCT_Y1CT	SEF_RHDO - N	SEF_RHST - N	SEF_RHCP_Y3CP	SEF_RHLP - N	SEF_RHC_HMC	SEF_RHIA - Y	AVG_KA - [200 - 400]
SEF_RHQM_HRMQ	SEF_RHQM_HRMQ	SEF_RHCV - N	SEF_RHCT_Y1CT	SEF_RHDO - N	SEF_RHST - N	SEF_RHCP_Y3CP	SEF_RHLP - N	SEF_RHC_HMC	SEF_RHIA - N	AVG_KA - [200 - 400]
SEF_RHQM_HRMQ	SEF_RHQM_HRMQ	SEF_RHCV - N	SEF_RHCT_Y1CT	SEF_RHDO - N	SEF_RHST - N	SEF_RHCP_Y3CP	SEF_RHLP - N	SEF_RHC_HMC	SEF_RHIA - Y	AVG_KA - [200 - 400]
SEF_RHQM_HRMQ	SEF_RHQM_HRMQ	SEF_RHCV - N	SEF_RHCT_Y1CT	SEF_RHDO - N	SEF_RHST - N	SEF_RHCP_Y3CP	SEF_RHLP - N	SEF_RHC_HMC	SEF_RHIA - N	AVG_KA - [200 - 400]
SEF_RHQM_HRMQ	SEF_RHQM_HRMQ	SEF_RHCV - N	SEF_RHCT_Y1CT	SEF_RHDO - N	SEF_RHST - N	SEF_RHCP_Y3CP	SEF_RHLP - N	SEF_RHC_HMC	SEF_RHIA - N	AVG_KA - [400 - 600]
SEF_RHQM_HRMQ	SEF_RHQM_HRMQ	SEF_RHCV - N	SEF_RHCT_Y1CT	SEF_RHDO - N	SEF_RHST - N	SEF_RHCP_Y3CP	SEF_RHLP - N	SEF_RHC_HMC	SEF_RHIA - N	AVG_KA_NP_NULL
SEF_RHQM_HRMQ	SEF_RHQM_HRMQ	SEF_RHCV - N	SEF_RHCT_Y1CT	SEF_RHDO - N	SEF_RHST - N	SEF_RHCP_Y3CP	SEF_RHLP - N	SEF_RHC_Y1C	SEF_RHIA - Y	AVG_KA_NP_NULL

10/10
0:41
0
0
0
Ln 10, Col 21 Spaces: 4 UTF-8 ORCL SQL MSSQL 5,095,270 rows 00:05:50 ALMEDBADMIN@SQLX/EXPRESS - EnemInkedServerSQL

Fonte: Elaborado pelo autor (2021)

A partir desta modelagem realizada, foram gerados os CSVs a partir da consulta de todos os registros gerados com as colunas específicas de cada grupo de análise para ser minerado, que são os casos de uso para avaliação.

3.3.6 Grupos de análise - Casos de uso





Foram divididos grupos de dados baseados nas tabelas originárias das colunas correspondentes a eles e cada grupo foi associado com colunas de média das notas de todas as áreas de conhecimento e a nota da redação, que correspondem a tabela de Fatos.

Os grupos de análise são informações da prova, com os dados provenientes da coluna Dimension_Info_Prova, que contém as informações sobre a prova junto a dados do município de residência da tabela de Informação pessoal, Localização, que possui os dados da nota junto dos dados de localidade de residência e de realização da prova, Dados Pessoais, retirados da tabela Dimension_Info_Pessoal, Situação do Ensino médio, retirada da tabela Dimension_Info_STEM e Questionários socioeconômicos, retirados da tabela Dimension_QSE. Este último grupo, por ser o maior, foi dividido em 4 subgrupos, sendo eles:

- Questionários de 1 a 6, que agrupam a situação familiar como escolaridade dos pais, renda mensal familiar e quantidade de pessoas na casa
- Questionários de 7 a 11, que agrupam a situação residencial como cômodos e veículos
- Questionários de 12 a 18, que agrupam os eletrodomésticos da residência

- Questionários de 19 a 25, que agrupam os aparelhos de multimídia e telecomunicações, como TV a cores, celular e internet

Figura 55 - Documentos CSVs de consultas para os casos de uso

Nome	Status	Tamanho	Tipo
 ExamFact_ExamInfo	✓	512.218 KB	Arquivo de Valores Separados por Vírgulas do Microsoft Excel
 ExamFact_Location	✓	525.253 KB	Arquivo de Valores Separados por Vírgulas do Microsoft Excel
 ExamFact_PersonalInfo	✓	656.036 KB	Arquivo de Valores Separados por Vírgulas do Microsoft Excel
 ExamFact_STEM	✓	748.061 KB	Arquivo de Valores Separados por Vírgulas do Microsoft Excel
 ExamFactSEF_FamilySituation	✓	809.609 KB	Arquivo de Valores Separados por Vírgulas do Microsoft Excel
 ExamFactSEF_HouseHoldAppliance	✓	733.681 KB	Arquivo de Valores Separados por Vírgulas do Microsoft Excel
 ExamFactSEF_HouseSituation	✓	585.919 KB	Arquivo de Valores Separados por Vírgulas do Microsoft Excel
 ExamFactSEF_MultiMediaTIC	✓	720.878 KB	Arquivo de Valores Separados por Vírgulas do Microsoft Excel

Fonte: elaborada pelo autor (2021)

A partir de cada avaliação em comparação com a nota, é possível associar cada um dos fatores com o desempenho do aluno. Nas informações pessoais, apenas foi analisada a localidade por UF, sendo ignorado o município, que foi tratado apenas no grupo de Localização, para identificar o perfil de realização do exame de acordo com as localidades residencial e de realização do exame.

Além destes grupos, também foi criado o grupo de perfil socioeconômico, que agrupa os dados dos questionários socioeconômicos com as informações pessoais dos alunos. O objetivo deste grupo é perfilar os participantes de acordo com suas características e sua situação social e econômica e a verificar como estes fatores se correlacionam com seu desempenho e presença nas provas.

3.3.6.1 Cenário de análise

Os Foi projetado um cenário de análise com parâmetro de confiança mínima de 80%, para que fossem selecionadas apenas as regras de associação com probabilidade significativa, para todos os casos de uso avaliados.

Os parâmetros de suporte foram variados de acordo com o tamanho de itens, ou colunas, no arquivo de transações gerado no R. Para Casos de uso com muitos itens em colunas, utilizou-se suporte de 1 a 10%, nos casos de poucos itens, usou o limiar 20% de suporte. Isso se deve ao fato de que com muitos itens, sua frequência se dilui e tende a ser mais baixa pela alta disponibilidade. Isso faz com que itens pouco frequentes, como por exemplo, níveis de escolaridade altos, tendem a ser excluídos da varredura.

Os valores de lift indicam o grau de correlação entre os conjuntos associativos, caso seja maior que 1 então a maior presença do antecedente gera uma ocorrência mais frequente de seu consequente, já se for abaixo de 0, quando maior a ocorrência antecedente, menos o consequente tende a ocorrer, gerando um impacto significativo que caracteriza os conjuntos como substitutivos. Caso o lift seja 1, então o antecessor gera seu sucessor em uma mesma proporção. As variações do lift apontam a graduação da associação, que expressa, basicamente, se há correlação negativa (lift<0) e positiva (lift>=1).

As regras buscadas possuem tamanho dos conjuntos de 1 a 4 itens, para selecionar conjuntos de múltiplos fatores com quantidade controlada. O limiar de alavancagem (Lift) arbitrado para a avaliação é de lift>1.05, desta forma pode-se selecionar apenas as regras de associação que têm um grau de correlação razoavelmente acima da relação direta quantitativa.

Figura 56 - Leitura e geração de regras dos casos de uso de questionários socioeconômicos

```
1 require(arules)
2 library(arules)
3
4 SEEP_FS=read.transactions("~/PathDirectory/TCC/Bases/Transformed/ExamFactSEF_FamilySituation.csv",format='basket',sep=",")
5 SEEP_FS<-SEEP_FS[-1,]
6 Arule_SEEP_FS<-apriori(SEEP_FS,parameter = list(support=0.05,confidence=0.80))
7 visual_SEEP_FS=data.frame(Antecedent=labels(lhs(Arule_SEEP_FS)),Consequent=labels(rhs(Arule_SEEP_FS)),Arule_SEEP_FS@quality)
8 RuleSEEP_FS_LS<-apriori(SEEP_FS,parameter = list(support=0.05,confidence=0.8,minlen=1,maxlen=4))
9 VisualdfSEEP_FS_LS=data.frame(Antecedent=labels(lhs(RuleSEEP_FS_LS)),Consequent=labels(rhs(RuleSEEP_FS_LS)),RuleSEEP_FS_LS@quality)
10
11 SEEP_MTIC=read.transactions("~/PathDirectory/TCC/Bases/Transformed/ExamFactSEF_MultiMediaTIC.csv",format='basket',sep=",")
12 SEEP_MTIC<-SEEP_MTIC[-1,]
13 RuleSEEP_MTIC<-apriori(SEEP_MTIC,parameter = list(support=0.1,confidence=0.8,minlen=1))
14 VisualdfSEEP_MTIC=data.frame(Antecedent=labels(lhs(RuleSEEP_MTIC)),Consequent=labels(rhs(RuleSEEP_MTIC)),RuleSEEP_MTIC@quality)
15
16 SEEP_HS=read.transactions("~/PathDirectory/TCC/Bases/Transformed/ExamFactSEF_HouseSituation.csv",format='basket',sep=",")
17 SEEP_HS<-SEEP_HS[-1,]
18 RuleSEEP_HS<-apriori(SEEP_HS,parameter = list(support=0.05,confidence=0.8,minlen=1,maxlen=4))
19 VisualdfSEEP_HS=data.frame(Antecedent=labels(lhs(RuleSEEP_HS)),Consequent=labels(rhs(RuleSEEP_HS)),RuleSEEP_HS@quality)
20
21 SEEP_HH=read.transactions("~/PathDirectory/TCC/Bases/Transformed/ExamFactSEF_HouseHoldAppliance.csv",format='basket',sep=",")
22 SEEP_HH<-SEEP_HH[-1,]
23 Rule_SEEP_HH<-apriori(SEEP_HH,parameter = list(support=0.1,confidence=0.80,minlen=1,maxlen=4))
24 visual_SEEP_HH=data.frame(Antecedent=labels(lhs(Rule_SEEP_HH)),Consequent=labels(rhs(Rule_SEEP_HH)),Rule_SEEP_HH@quality)
```

Fonte: elaborada pelo autor (2021)

Figura 57 - Leitura e geração de regras dos casos de uso de informações pessoais, da prova, do ensino médio e de localidade

```
1 require(arules)
2 library(arules)
3
4 EFPI=read.transactions("~/PathDirectory/TCC/Bases/Transformed/ExamFact_PersonalInfo.csv",format='basket',sep=",")
5 EFPI<-EFPI[-1,]
6 RuleEFPI<-apriori(EFPI,parameter = list(support=0.01,confidence=0.8,minlen=1,maxlen=4))
7 VisualdfEFPI=data.frame(Antecedent=labels(lhs(RuleEFPI)),Consequent=labels(rhs(RuleEFPI)),RuleEFPI@quality)
8
9 EFEL=read.transactions("~/PathDirectory/TCC/Bases/Transformed/ExamFact_ExamInfo.csv",format='basket',sep=",")
10 EFEL<-EFEL[-1,]
11 RuleEFEL<-apriori(EFEL<-EFEL[-1,],parameter = list(support=0.05,confidence=0.8,minlen=1,maxlen=4))
12 VisualdfEFEL=data.frame(Antecedent=labels(lhs(RuleEFEL)),Consequent=labels(rhs(RuleEFEL)),RuleEFEL@quality)
13
14 STEM=read.transactions("~/PathDirectory/TCC/Bases/Transformed/ExamFact_STEM.csv",format='basket',sep=",")
15 STEM<-STEM[-1,]
16 RuleSTEM<-apriori(STEM,parameter = list(support=0.05,confidence=0.8,minlen=1,maxlen=4))
17 VisualdfSTEM=data.frame(Antecedent=labels(lhs(RuleSTEM)),Consequent=labels(rhs(RuleSTEM)),RuleSTEM@quality)
18
19 EFL=read.transactions("~/PathDirectory/TCC/Bases/Transformed/ExamFact_Location.csv",format='basket',sep=",")
20 EFL<-EFL[-1,]
21 RuleEFL<-apriori(EFL,parameter = list(support=0.01,confidence=0.8,minlen=1,maxlen=4))
22 VisualdfEFL=data.frame(Antecedent=labels(lhs(RuleEFL)),Consequent=labels(rhs(RuleEFL)),RuleEFL@quality)
```

Fonte: elaborada pelo autor (2021)

3.3.7 Análise e assimilação de resultados

Nesta etapa do KDD, é realizado o levantamento geral de todos os resultados coletados e são levantadas hipóteses que o expliquem. Neste contexto, tal como a análise de dados ad-hoc, é realizada a abordagem de estatística descritiva para levantar os dados mais relevantes coletados. Os dados foram analisados nos dataframes gerados nos scripts das figuras (acima) e (acima), que agrupam as informações das regras de associação mineradas.

Para extrair as regras de associação úteis, além dos limiares de controle avaliados, foi aplicado filtro dos itens consequentes correspondente a variáveis significativas no impacto do inscrito nas provas, como o prefixo “{AVG_KA -”, para filtrar todas as consequências relacionadas a média nas áreas de conhecimento, do inglês *Average_KnowledgeAreas*, , que correspondem às provas objetivas, e o prefixo “{ETGI -”, para levantar consequentes de intervalos de notas da redação, do inglês *EssayTotalGradesIntervals*. Estes filtros também foram aplicados aos antecedentes, para aplicar a análise reversa, ou seja, o fato de o aluno ter determinado desempenho implicar em determinada característica pessoal, social ou econômica.

3.3.7.1 Análise dos agrupamentos – casos de uso

Os levantamentos das regras de associação com confiança superior a 80%, suporte mínimo de 0,5%, 1%, 5%, ou 10% foram agrupados em casos de uso que correspondem aos seus tipos de dados e suas relações com as métricas de desempenho na redação e nas demais provas. Nesta seção, todas estas regras serão detalhadas de acordo com cada um dos casos.

3.3.7.2 Avaliação dos resultados

Através da mineração de dados, é possível identificar que os fatores mais impactantes no desempenho do participante são a sua situação familiar, seguido de características pessoais como raça e idade. Também pode-se evidenciar que certas localidades costumam ser associadas ao baixo desempenho, como os estados de PA e BA e outros com desempenho mais elevado, como SP e MG.

Analisando a situação dos eletrodomésticos, não há grande correlação deste grupo em específico com o rendimento dos participantes, apenas verificou-se que pessoas com apenas um refrigerador, sem máquina de lavar roupas e nem máquina de lavar louças possui uma chance elevada de faltar a redação e a uma prova objetiva, o que pode evidenciar um perfil de vulnerabilidade social que influencie na condição de estar presente no dia do exame. Aqui,

todas as regras possuem alavancagem e confiança elevados, o que pode apontar um perfil específico.

Sobre os aparelhos de multimídia e telecomunicação, pouco tendem a afetar ou ter relação com desempenho, apesar de certos padrões serem frequentes tanto se associados a notas baixas ou alta, como não possuir telefone fixo ou TV por assinatura. A única relação significativa é o acesso à internet e a média das objetivas entre 400 e 600, que é bem frequente.

A respeito da prova, não se detectou uma relação tão expressiva de algum item específico correlato as métricas de desempenho, todas com alavancagem sempre próxima a 1. As únicas associações significativas avaliadas são a do antecessor ser a nota da redação acima de 750 e o consequente impactado ser a média das notas das áreas de conhecimento ser de 400 a 600 e as maiores notas de redação e prova objetiva ocorrerem em conjunto com a escolha da língua inglesa e vice-versa, nos casos em que a redação não é zerada. Quase todas as regras apontam a redação normalizada, sem zeramento, portanto não há nenhuma referência causal relevante entre as características das provas realizadas.

Ao avaliar a situação doméstica da residência, não há muitas associações relevantes detectáveis, apenas correlações entre desempenho baixo nas médias das objetivas e redação em conjunto com não existência de veículos com o fato de existir somente um banheiro na casa. Outro antecessor que costuma aparecer é o de possuir 1 e também 2 quartos, todas essas regras englobam baixo desempenho e têm como consequência apenas um banheiro.

Nas informações pessoais, encontra-se forte correlação entre os antecessores de idade entre 18 e 24 anos e nacionalidade nativo brasileira e a falta em alguma área de competência. Este perfil social compreende aos jovens recém-saídos em ensino médio em sua maioria, que, ingressam no mercado de trabalho e ou na universidade, o que pode ocasionar desistência. Também foi detectado uma ocorrência frequente dos conjuntos de antecessores contendo nota na redação acima de 750, tipo racial branco e UF de residência ser SP, e estes conjuntos, tal como a presença do item idade inferior a 18 anos, sexo masculino e estado civil solteiro, e todos estes itens fortemente associados a consequência de média nas provas objetivas entre 400 e 600. Já um conjunto específico composto por nota na redação entre 250 e 500, sexo feminino e a nacionalidade brasileira nativa tende a possuir o consequente de nota nas áreas de conhecimento de 200 a 400 com confiança e alavancagem consideráveis.

Na situação de ensino médio, bastante regras aproveitáveis foram extraídas a partir das medidas de interesse com limiares de alto padrão. Verificou-se que antecessores compostos

por média das provas objetivas entre 200 e 400, localidade da escola urbana e ou a escola ser pública, também inferem na consequência desta escola ser estadual.

Também se verifica que alunos de escola pública com esta nota costumeiramente não são treineiros, o que evidencia que mesmo alunos formados ou prestes a se formar deste tipo de instituição costumam ter baixos rendimentos. O mesmo não se verifica na redação, onde antecedentes que contém nota acima de 750 e entre 500 e 750 ocorrem com frequência considerável. Assim, pode-se aferir tendência de baixos desempenhos dos alunos de escola pública estadual nas objetivas, mas não se pode afirmar isto para a redação.

Em se tratando da situação familiar do participante, a maioria dos itens antecedentes são frequentes e tendem a ocasionar umas às outras se agrupados com as notas em conjuntos antecedentes, o que aponta fortes correlações com potencial de causalidade elevado.

Os itens que mais tendem a impactar na consequência de médias entre 400 e 600 são o número de pessoas na casa ser de 1 a 4, grupo ocupacional dos pais ser profissional liberal ou empreendedor, e a média na redação ser acima de 750. Fatores fortemente ligados a falta em uma das áreas de conhecimento são os antecedentes compostos por falta na redação, grupo ocupacional do pai ser primeiro setor e escolaridade deste ser ensino fundamental I incompleto.

O conjunto de itens composto por médias das objetivas entre 200 e 400, o pai ser do primeiro setor e a renda mensal ser nula também implicam na mãe ser do mesmo grupo ocupacional, consequentemente. Foi analisado também um conjunto em que a nota da redação é de 250 a 500, a quantidade de pessoas na casa é de 5 a 8 e a mãe pertence ao primeiro setor, e seu consequente é a média das objetivas estar entre 200 e 400, com uma confiança e alavancagem relativamente acima do mínimo definido.

Desta forma, pode-se inferir que o grupo ocupacional dos pais, sobretudo da mãe, tem impacto significativo na presença do participante e em seu desempenho, bem como os níveis de renda. No mais, tudo isto em patamares similares, a quantidade de pessoas em casa, caso seja acima de quatro, tende a ser o diferencial que ocasione impactos negativos no rendimento em provas objetivas.

3.4 ANÁLISE DE DADOS A PARTIR DO CONJUNTO DE DADOS MODELADOS

Nesta seção, é explicitado todo o processo de análise de dados ad-hoc de caráter descritivo da base de dados.

3.4.1 ETL

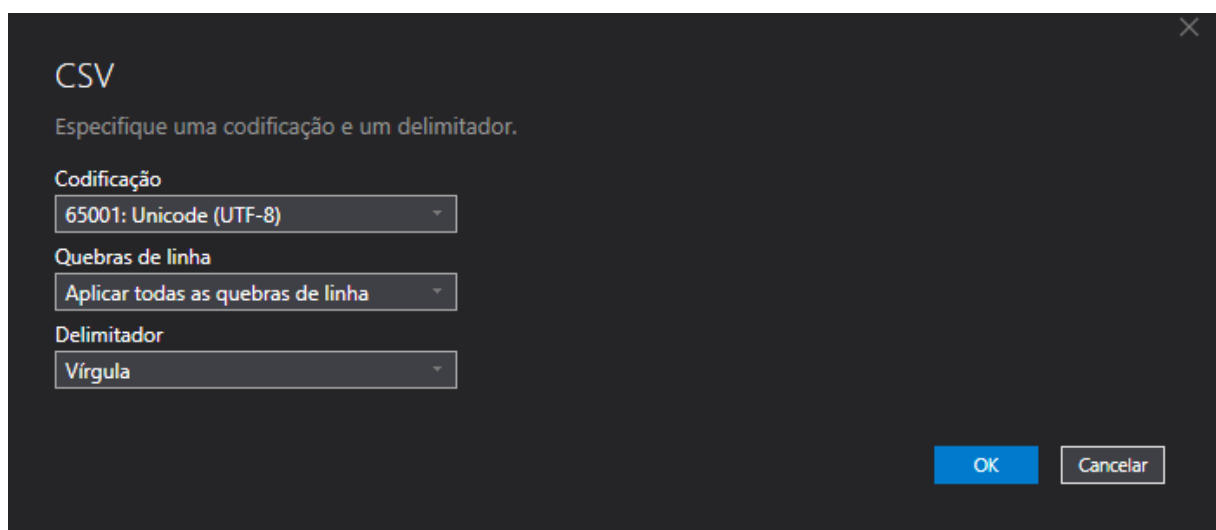
Os dados foram extraídos do Azure Data Lake, conectados ao Visual Studio, para poder tratar os dados contidos nas tabelas, para poder conectar ao Azure Analysis Service e gerar o visual no Power BI.

Todas as transformações realizadas foram realizadas dentro do Power Query do Visual Studio

3.4.1.1 Ajustes da importação

Todas as tabelas importadas do Azure Data Lake, sofreram alguns ajustes para poder se adequar as configurações de visualização do Visual Studio.

Figura 58 - Alteração da codificação das tabelas



Fonte: Elaborado pelo ato através do Visual Studio

Conforme na figura 58 a alterações foi na codificação que foi eleito o alfabeto de Unicode UTF-8, as outras opções de quebra de linhas e delimitador já estavam como padrão.

Logo após isso, a primeira linha das tabelas foi promovida a cabeçalhos das colunas.

3.4.1.2 Transformação das tabelas

Os dados das tabelas originais estão em formato de índice, os dados estão em forma de número ou letra, exceto as colunas que se referem a localidade.

Para que possamos ter uma melhor visibilidade dos dados as colunas originais foram utilizadas para gerarem uma nova coluna a partir de seus dados, como mostra a figura abaixo

Figura 59 - Transformação das colunas sobre o a nacionalidade dos participantes.

123 TP_NACIONALIDADE	ABC 123 NACIONALIDADE
4	Brasileiro(a) Nato(a), nascido(a) no exterior
1	Brasileiro(a)
1	Brasileiro(a)
1	Brasileiro(a)
1	Brasileiro(a)
1	Brasileiro(a)
1	Brasileiro(a)
1	Brasileiro(a)
1	Brasileiro(a)
2	Brasileiro(a) Naturalizado(a)
1	Brasileiro(a)

Fonte: Elaborado pelo ato através do Visual Studio

Na figura 59 é demonstrado como foram geradas as colunas, através das associações dos dados contidos em cada coluna, para fazer essa associação foi necessário consultar

Figura 60 - Adicionando uma coluna condicional

Adicionar Coluna Condicional

Adicionar uma coluna condicional que é calculada das outras colunas ou valores.

Nome da nova coluna

NACIONALIDADE

	Nome da Coluna	Operador	Valor		Saída
Se	TP_NACIONALIDA...	igual a	ABC 123 0	Então	ABC 123 Não informado
Senã...	TP_NACIONALIDA...	igual a	ABC 123 1	Então	ABC 123 Brasileiro(a)
Senã...	TP_NACIONALIDA...	igual a	ABC 123 2	Então	ABC 123 Brasileiro(a) Naturalizado(a)
Senã...	TP_NACIONALIDA...	igual a	ABC 123 3	Então	ABC 123 Estrangeiro(a)
Senã...	TP_NACIONALIDA...	igual a	ABC 123 4	Então	ABC 123 Brasileiro(a) Nato(a), nascido(a)

Adicionar Cláusula

Senão

ABC 123 null

OK

Cancelar

Fonte: Elaborado pelo ato através do Visual Studio

As colunas que foram utilizadas como base para adição da nova coluna foram removidas

Além deste tipo de transformação, também houve a criação de faixas de conjunto de dados, com o intuito de agrupar os dados e criar uma visualização de conjuntos como o exemplo da figura 61, essas colunas que foram utilizadas para gerar essas faixas não foram excluídas pois elas serão utilizadas na visualização do Power BI.

Figura 61 - Distribuição por faixa da tabela Info_Pessoal

1 ² 3 NU_IDADE	ABC 123 FX_IDADE
21	18 à 24 anos
16	Abaixo dos 18 anos
18	Abaixo dos 18 anos
23	18 à 24 anos
23	18 à 24 anos
31	31 à 36 anos
30	25 à 30 anos
26	25 à 30 anos
19	18 à 24 anos
17	Abaixo dos 18 anos

Fonte: Elaborado pelo ato através do Visual Studio

A configuração utilizada para essa distribuição foi a coluna condicional como está na figura 62 abaixo

Figura 62 - Criação da coluna condicional para a faixa de idade na tabela Info_Pessoal

Adicionar Coluna Condicional

Adicionar uma coluna condicional que é calculada das outras colunas ou valores.

Nome da nova coluna

FX_IDADE

	Nome da Coluna	Operador	Valor		Saída
Se	NU_IDADE	é menor que ou i...	18	Então	Abaixo dos 18 anos
Senã...	NU_IDADE	é menor que ou i...	24	Então	18 à 24 anos
Senã...	NU_IDADE	é menor que ou i...	30	Então	25 à 30 anos
Senã...	NU_IDADE	é menor que ou i...	36	Então	31 à 36 anos
Senã...	NU_IDADE	é menor que ou i...	42	Então	37 à 42 anos
Senã...	NU_IDADE	é maior que ou ig...	43	Então	Maior que 43 anos

Adicionar Cláusula

Senão null

OK

Cancelar

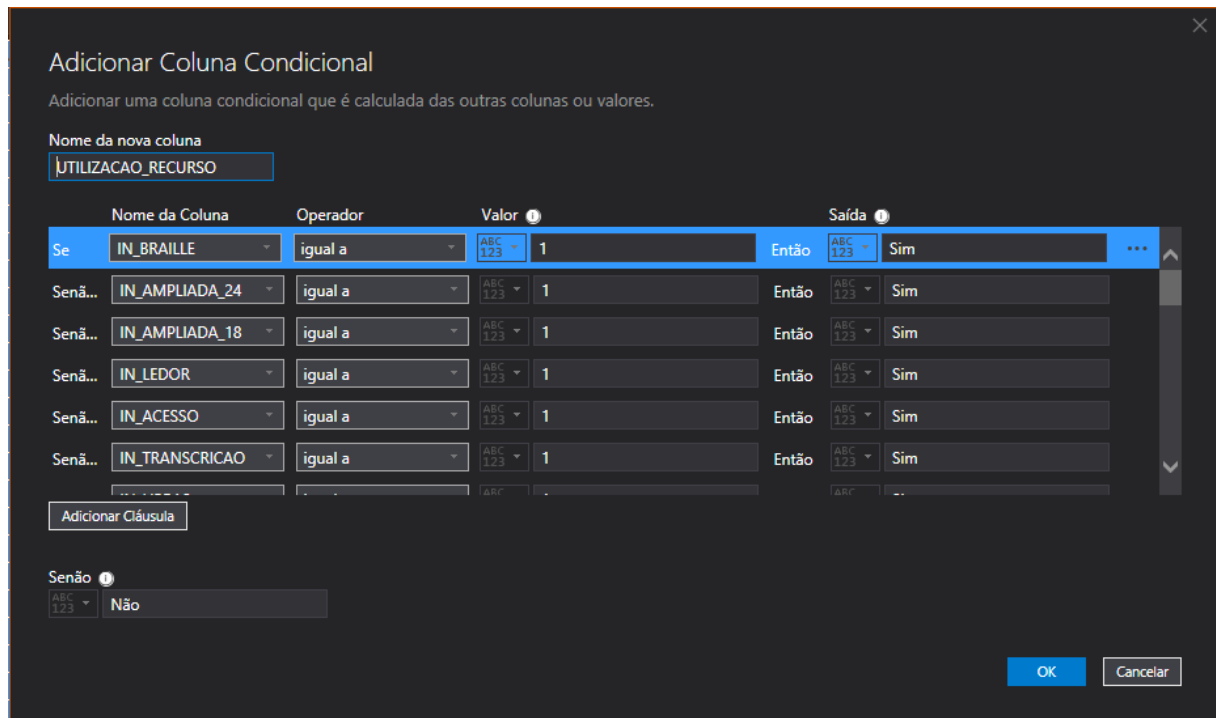
Fonte: Elaborado pelo ato através do Visual Studio

Algumas colunas foram resumidas como as colunas da tabela de inclusão, pois era esse estudo era necessário saber somente se os participantes solicitaram atendimento específico, especializado ou solicitaram a utilização de algum recurso para realização do exame. No dicionário disponibilizado pelo INEP há 13 colunas a respeito da solicitação de atendimento especializado, 4 colunas sobre a solicitação de atendimento específico e 34 colunas sobre a solicitação de recurso, no caso dessa tabela em específico essas colunas foram resumidas de 51 colunas para 4 colunas como demonstrado na figura 63.

[illegible]

O processo utilizado para a geração das colunas resumidas foi o de colunas condicionais, como descrito na figura 64.

Figura 64 - Adicionando coluna condicional a partir das colunas sobre a utilização de recursos



Fonte: Elaborado pelo ato através do Visual Studio

3.4.2 Estatística gerais

Nesta seção nós buscamos exibir as informações mais abrangente possível, já que estamos tratando de uma base dados muito grandes, nela podemos encontrar muitas informações que podem ser valiosas.

3.4.2.1 Estatísticas de desempenho

Como está na figura 65 buscamos mostrar exatamente as informações que correspondem a nota, como a quantidade de inscritos que atingiu uma faixa das notas, para podermos ter ciência onde há maior concentração de inscritos por conjunto faixas de nota.

Figura 65 - quantidade de inscritos por faixa da média nota da prova objetiva

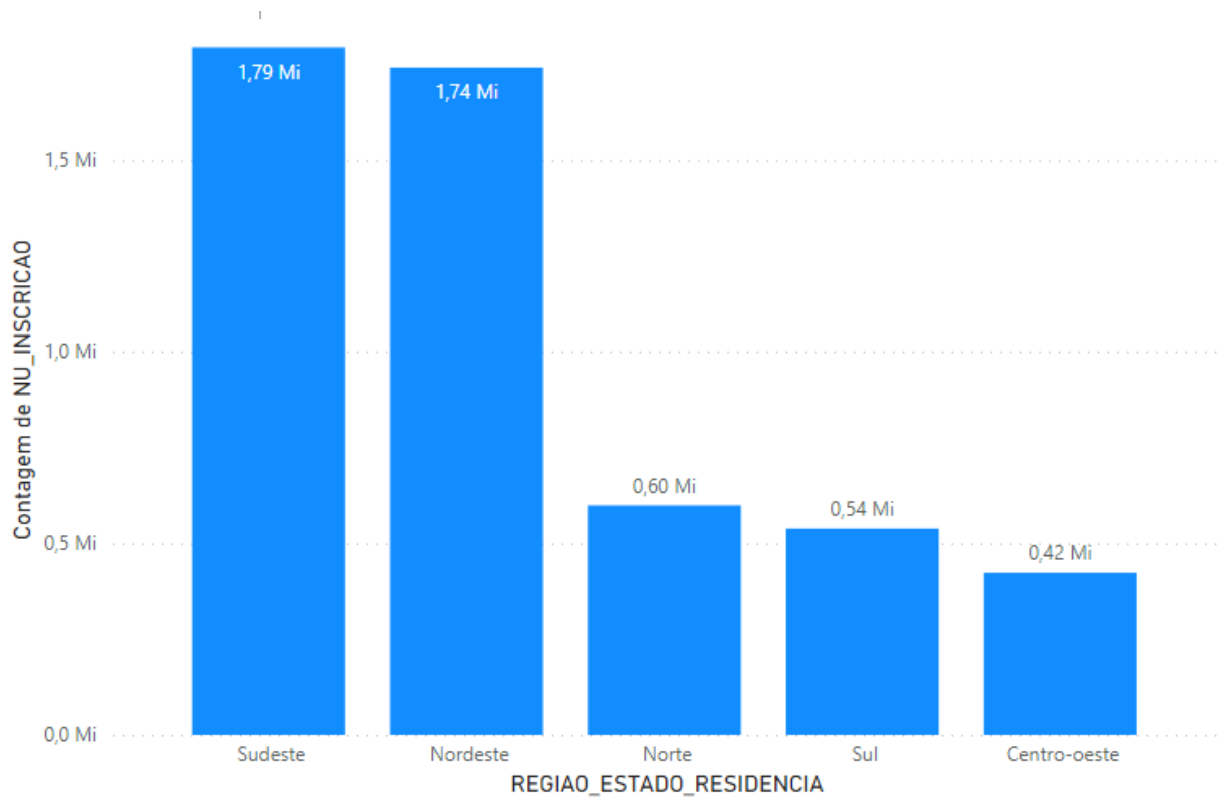
FX_MED_NOTA_PROVA_OBJ	Contagem de NU_INSCRICAO
	1163700
200 a 400 pontos	116496
400 a 600 pontos	3363075
Até 200 pontos	2863
Maior de 600 pontos	449136
Total	5095270

Fonte: Elaborado pelo autor (2021)

3.4.2.2 Estatísticas das informações pessoais e geográficas

Conforme a figura 66 nas informações pessoais exibimos as quantidades das características dos inscritos.

Figura 66 - Gráfico de colunas onde exhibe a quantidade de inscritos por regiões do Brasil, situando as suas respectivas residências

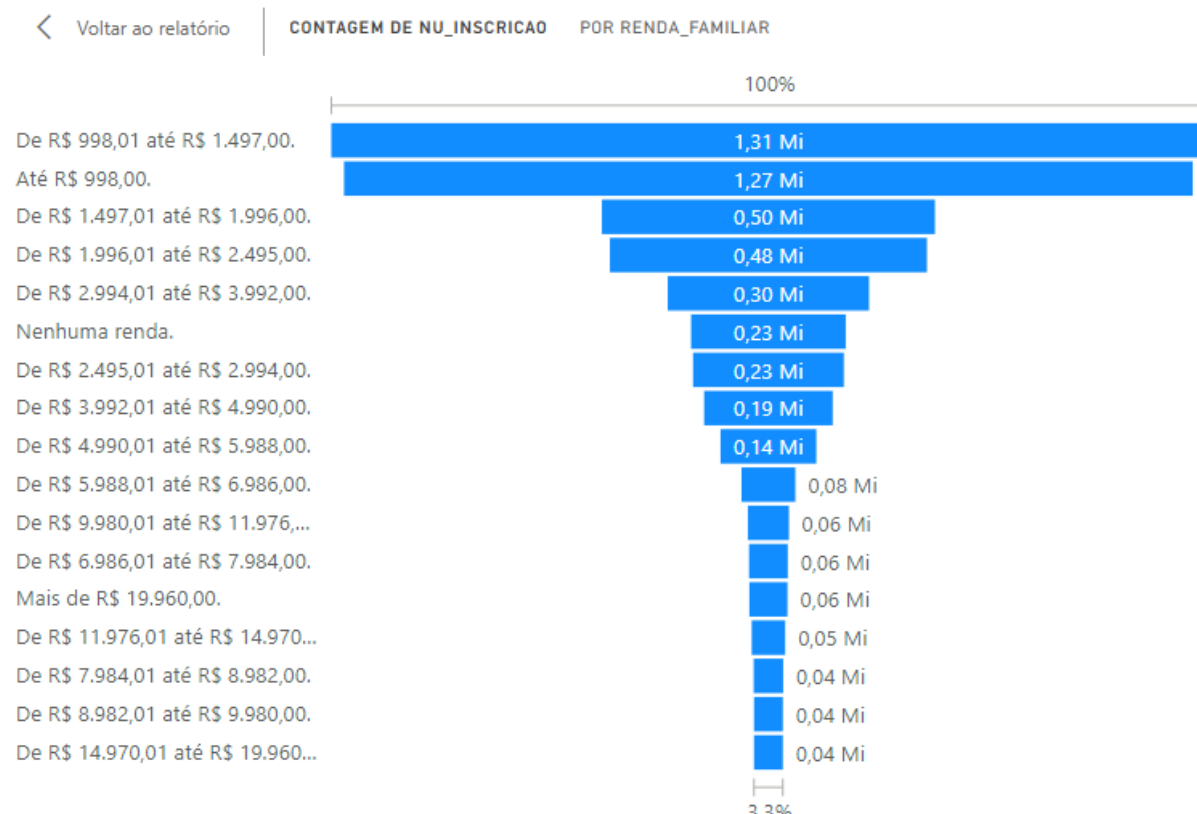


Fonte: Elaborado pelo autor (2021)

3.4.2.3 Estatísticas de socioeconômicas

Como está na figura 67 pode ser visto quantidade de inscritos conforme as informações sobre da renda familiar.

Figura 67 - Distribuição da renda familiar pela quantidade de inscritos

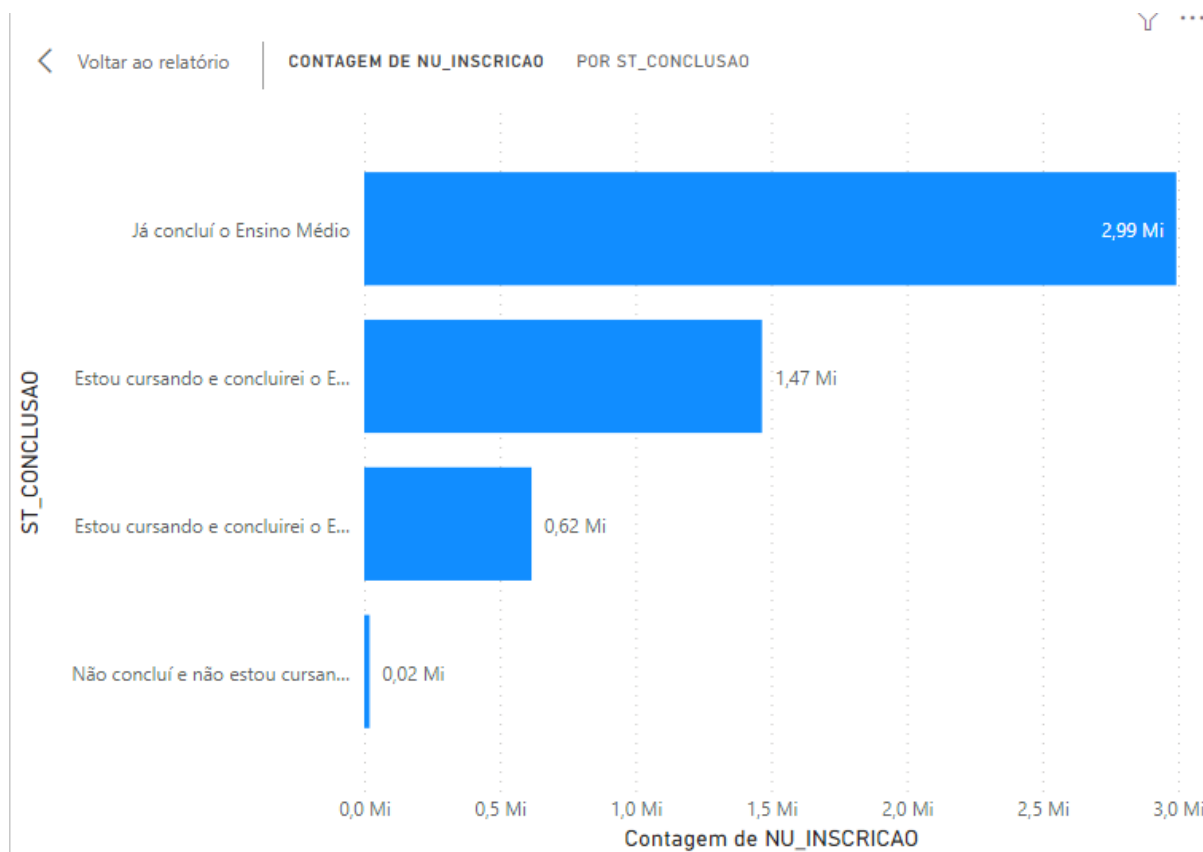


Fonte: Elaborado pelo autor

3.4.2.4 Estatísticas da situação de conclusão do ensino médio

Sobre a situação de conclusão do ensino médio dos inscritos nós também levantamos dados simples para ter noção da quantidade dos conjuntos de dados, como está na figura 68.

Figura 68 - Gráfico de barras que mostra a comparação dos inscritos com a situação do ensino médio

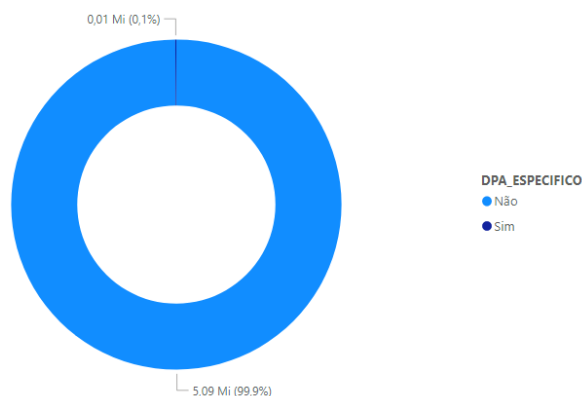


Fonte: Elaborado pelo autor (2021)

3.4.2.5 Estatísticas de inclusão

No gráfico da figura 69 é visto a diferença dos inscritos que solicitaram um atendimento específico e o que não solicitaram.

Figura 69 - Gráfico de rosca exibindo a diferença de quantidade de inscritos que solicitaram atendimento específico



Fonte: Elaborado pelo autor

4. RESULTADOS COMPARATIVOS E ANÁLISE DIAGNÓSTICA

Neste capítulo, são expostos os resultados explícitos dos dados coletados a partir do levantamento estatístico das consultas ad-hoc e das hipóteses descobertas da base através da aplicação do algoritmo automatizado de mineração de dados, que configuram a análise estatística descritiva. Além disto, é referenciado uma interpretação especializada pautada no embasamento analítico de profissionais da área da educação e assistência social, para respaldar o diagnóstico dos resultados e, assim, aplicar a abordagem de análise diagnóstica destes dados.

4.1 ESTATÍSTICAS DO ENEM

Nesta seção abordaremos a capacidade de exploração das informações através dos gráficos do Power BI em um caso específico. Utilizamos o gráfico de Pareto e o de árvore hierárquica. Na figura 70 há um gráfico de árvore hierárquica e podemos analisar que a média da redação vai se adequando em todo o caminho percorrido, no final desse caminho pode-se concluir que o participante que ficou com nota na faixa de 500 a 750 pontos na redação, realizou a prova no Sudeste e teve uma nota da prova objetiva na faixa de 400 a 600 pontos, possui uma média de pontos na redação de 607,44 no exame daquele respectivo ano. Entretanto os participantes que tiveram os mesmos atributos, mas na nota da prova objetiva obtiveram uma nota maior que 600 pontos, tiveram uma nota maior na redação, comparado o dado anterior.

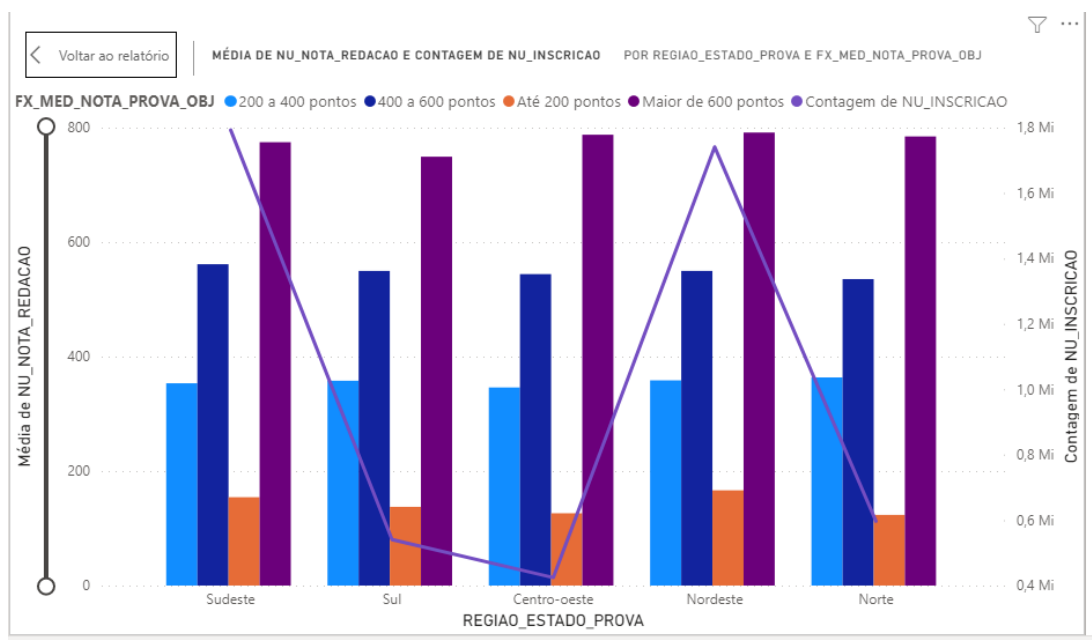
Figura 70 - Gráfico de árvore hierárquica exibindo o desenvolvimento de um participante através de seus atributos



Fonte: Elaborado pelo autor (2021)

Já no gráfico de Pareto que está na figura 71, podemos observar através da linha que corta o gráfico que a concentração de quantidade de participantes está nas regiões do nordeste e do sudeste e comparando com as outras regiões a quantidade é bem maior. Além disso podemos ressaltar que a maior média nota da redação persiste na mesma faixa dos maiores de 600 pontos da prova objetiva, mas neste caso não se aplica só ao Sudeste, como na figura 70 acima, mas a todas as regiões do Brasil. Entre as outras faixas de notas da prova objetiva a média da nota de redação é qual é o mesmo valor. O que podemos concluir é que só as regiões em si do Brasil não influenciam na média da nota dos participantes, mas se agregamos a outros valores, certamente haverá alguma diferença

Figura 71 - Gráfico de Pareto, para comparar as regiões e as notas dos participantes



Fonte: Elaborado pelo autor (2021)

4.2 MINERAÇÃO DE DADOS

Nesta seção são detalhados os limiares analisados na coleta das regras de associação descobertas e os parâmetros de cada uma delas, dado que foram selecionadas pela não trivialidade total e pelo nível de Lift superior a 1.05, desta forma são consideradas apenas as associações positivas, isto é, quando um conjunto de itens impacta positivamente na consequência e não o contrário.

Figura 72 -: Caso de uso - Informações pessoais

ANTECEDENTE(S)	CONSEQUENTE	FREQUÊNCIA	CONFIANÇA	ALAVANCAGEM
[Falta na redação] [Estado civil: solteiro] [Faixa etária: de 18 a 24 anos]	Falta em uma das áreas de competência	12%	100%	3,65
[Falta na redação] [Nacionalidade: Nativo Brasileiro] [Faixa etária: de 18 a 24 anos]	Falta em uma das áreas de competência	12,90%	100%	3,65
[Média total das notas: entre 400 e 600] [Nota na redação: entre 500 e 750] [Faixa etária: de 18 a 24 anos]	Estado civil solteiro	10,30%	94,90%	1,09
[Média total das notas: entre 200 e 400] [Faixa etária: entre 18 e 24 anos] [Tipo racial: pardo ou mestiço]	Estado civil solteiro	10,90%	94,10%	1,09

Fonte: elaborada pelo autor (2021)

Figura 73 - Caso de uso - Informações pessoais - 2

ANTECEDENTE(S)	CONSEQUENTE	FREQUÊNCIA	CONFIANÇA	ALAVANCAGEM
[Nota na redação: entre 750 e 1000] [Nacionalidade: nativo brasileiro]	Média total das notas entre 400 e 600	10,30%	88,20%	2,5
[Nota na redação: entre 750 e 1000] [UF de residência: SP] [Tipo racial: branco]	Média total das notas entre 400 e 600	1,30%	94,60%	2,69
[Nota na redação: entre 750 e 1000] [Tipo racial: branco] [Sexo: Masculino]	Média total das notas entre 400 e 600	2%	94,10%	2,67
[Nota na redação: entre 750 e 1000] [UF de residência: SP] [Estado Civil: Solteiro]	Média total das notas entre 400 e 600	1,70%	93,50%	2,65
[Nota na redação: entre 750 e 1000] [Tipo racial: branco] [Idade: Menor de 18 anos]	Média total das notas entre 400 e 600	2,20%	92,70%	2,63
[Nota na redação: entre 250 e 500] [Sexo: Feminino] [Nacionalidade: nativo Brasileiro]	Média total das notas entre 200 e 400	8,70%	87,50%	2,14

Fonte: elaborada pelo autor (2021)

Figura 74 - Caso de uso - Situação familiar - 1

ANTECEDENTE(S)	CONSEQUENTE	FREQUÊNCIA	CONFIANÇA
[Nota da redação: maior que 750] [Pessoas na casa: de 1 a 4]	Média das áreas de conhecimento entre 400 e 600	8%	89%
[Falta na redação] [Pessoas na casa: de 1 a 4 na casa] [Grupo ocupacional da mãe: Terceiro Setor e Comércio]	Falta em uma(s) área(s) de conhecimento	8%	100%
[Falta na redação] [Grupo ocupacional do pai: Primeiro setor ou extrativista]	Falta em uma(s) área(s) de conhecimento	5,70%	100%
[Falta na redação] [Escolaridade do pai: Não concluiu o primeiro ciclo do fundamental]	Falta em uma(s) área(s) de conhecimento	6,40%	100%
[Nota da redação: entre 250 e 500] [Renda mensal familiar: abaixo de R\$ 998,00]	Média nas áreas de conhecimento entre 400 e 600	5,30%	83,70%

Fonte: elaborada pelo autor (2021)

Figura 75 - Caso de uso - Situação familiar - 2

ANTECEDENTE(S)	CONSEQUENTE	FREQUÊNCIA	CONFIANÇA	ALAVANCAGEM
[Média na redação: entre 750 e 1000] [Pessoas na casa: de 1 a 4]	Média nas áreas de conhecimento entre 400 e 600	8,20%	88,90%	2,53%
[Média nas áreas de conhecimento: entre 200 e 400] [Grupo ocupacional da mãe: Primeiro setor ou extrativista]	Grupo ocupacional do pai: Primeiro setor ou extrativista	7,30%	82,40%	3,84
[Média nas áreas de conhecimento: entre 200 e 400] [Grupo ocupacional da mãe: Primeiro setor ou extrativista] [Renda mensal familiar: Sem renda]	Grupo ocupacional do pai: Primeiro setor ou extrativista	1,10%	89,90%	4,19
[Média na redação: entre 750 e 1000] [Pessoas na casa: de 1 a 4 na casa] [Grupo ocupacional da mãe: Profissional liberal ou pequeno empresário(mais de 10 funcionários)]	Média nas áreas de conhecimento entre 400 e 600	1%	95,60%	2,72

Fonte: elaborada pelo autor (2021)

Figura 76 - Caso de uso - Situação familiar - 3

ANTECEDENTE(S)	CONSEQUENTE	FREQUÊNCIA	CONFIANÇA	ALAVANCAGEM
[Média na redação: entre 750 e 1000] [Pessoas na casa: de 1 a 4 na casa] [Grupo ocupacional do pai: Profissional liberal ou pequeno empresário(mais de 10 funcionários)]	Média nas áreas de conhecimento entre 400 e 600	1,50%	95,50%	2,71
[Nota da redação: entre 250 e 500] [Escolaridade do pai: não concluiu o primeiro ciclo do fundamental] [Escolaridade da mãe: não concluiu o primeiro ciclo do fundamental]	Média nas áreas de conhecimento entre 200 e 400	1,90%	80,60%	2,15
[Nota da redação: entre 250 e 500] [Pessoas na casa: de 1 a 4 na casa] [Renda mensal familiar: abaixo de R\$ 998,00]	Média nas áreas de conhecimento entre 200 e 400	3,50%	82,50%	2,2
[Nota da redação: entre 250 e 500] [Renda mensal familiar: sem renda]	Média nas áreas de conhecimento entre 200 e 400	1%	84,9	2,27
[Nota da redação: entre 250 e 500] [Pessoas na casa: de 5 a 8 na casa] [Grupo ocupacional da mãe: Primeiro setor ou extrativista]	Média nas áreas de conhecimento entre 200 e 400	1,30%	86,40%	2,3

Fonte: elaborada pelo autor (2021)

Figura 77 - Caso de uso - Situação doméstica

ANTECEDENTE(S)	CONSEQUENTE	FREQUÊNCIA	CONFIANÇA	ALAVANCAGEM
[Nota da redação: entre 250 e 500] [Residência não possui veículo - carro]	Residência possui um banheiro	11,30%	88,80%	1,26
[Nota da redação: entre 250 e 500] [Residência não possui veículo - Moto]	Residência possui um banheiro	11,80%	80,40%	1,26
[Nota da redação: entre 250 e 500] [Residência possui dois quartos]	Residência possui um banheiro	9,10%	87,30%	1,23
[Média total das notas: entre 200 e 400] [Nota da redação: entre 250 e 500]	Residência possui um banheiro	11,90%	81,20%	1,15
[Nota da redação: entre 500 e 750] [Residência não possui veículo - carro]	Residência possui um banheiro	19%	86,20%	1,22

Fonte: elaborada pelo autor (2021)

Figura 78 - Caso de uso - Multimídia e telecomunicação

ANTECEDENTE(S)	CONSEQUENTE	FREQUÊNCIA	CONFIANÇA	ALAVANCAGEM
[Média das notas entre 200 e 400] [Residência não possui computador] [Residência não possui telefone fixo]	Residência não possui TV por assinatura	23,10%	91,80%	1,17
[Média das notas entre 200 e 400] [Residência não possui computador]	Residência não possui telefone fixo	20,40%	91,80%	1,2
[Média das notas entre 200 e 400] [Residência possui uma TV colorida]	Residência não possui telefone fixo	25,20%	87,50%	1,14
[Média das notas entre 400 e 600]	Residência possui acesso a internet	31,20%	88,70%	1,14
[Nota na redação entre 500 e 750] [Residência possui uma TV colorida] [Residência não possui telefone fixo]	Residência não possui TV por assinatura	21,20%	89,60%	1,14

Fonte: elaborada pelo autor (2021)

Figura 79 - Caso de uso - Eletrodomésticos

ANTECEDENTE(S)	CONSEQUENTE	FREQUÊNCIA	CONFIANÇA	ALAVANCAGEM
[Falta na redação] [Não possui máquina lava-louças] [Residência possui um refrigerador]	Falta em uma das áreas de conhecimento	22%	99%	3,65
[Falta na redação] [Não possui máquina de lavar roupas] [Residência possui um refrigerador]	Falta em uma das áreas de conhecimento	20%	99%	3,65

Fonte: elaborada pelo autor (2021)

Figura 80 - Caso de uso - Situação do ensino médio

ANTECEDENTE(S)	CONSEQUENTE	FREQUÊNCIA	CONFIANÇA	ALAVANCAGEM
[Nota na redação: entre 500 e 750] [Tipo de dependência administrativa da escola: Pública]	Não é treineiro	11,40%	100%	1,13
[Média total das notas: entre 200 e 400] [Tipo de dependência administrativa da escola: Pública]	Não é treineiro	11,90%	100%	1,13
[Média total das notas: entre 200 e 400] [Tipo de dependência administrativa da escola: Pública]	Prestes a concluir o ensino médio	11,90%	100%	3,4
[Nota na redação: entre 500 e 750] [Tipo de dependência administrativa da escola: Pública]	Prestes a concluir o ensino médio	11,40%	100%	3,4
[Média total das notas: entre 200 e 400] [Nota na redação: entre 250 e 500] [Localização da escola: não se aplica, pois já concluiu o EM]	Tipo de dependência administrativa(escola) não aplicável, já concluiu o ensino médio	10,90%	100%	1,29
[Nota na redação: entre 500 e 750] [Localidade da escola: Urbana] [Situação operacional da escola: ativa]	Não é treineiro	10,10%	100%	1,13

Fonte: elaborada pelo autor (2021)

Figura 81 - Caso de uso - Situação do ensino médio - 2

ANTECEDENTE(S)	CONSEQUENTE	FREQUÊNCIA	CONFIANÇA	ALAVANCAGEM
[Média total das notas: entre 200 e 400] [Tipo de escola: Pública] [Localidade da escola: Urbana]	Dependência administrativa da escola é Estadual	8,90%	96%	5,31
[Média total das notas: entre 200 e 400] [Situação operacional da escola: ativa]	Tipo de escola é pública	9,50%	97%	3,94
[Média total das notas: entre 200 e 400] [Nota na redação: entre 500 e 750] [Situação do ensino médio: prestes a concluir]	Tipo de escola é pública	5,90%	93%	3,8
[Média total das notas: entre 200 e 400] [Localidade da escola: Urbana]	[Tipo de dependência administrativa da escola: Estadual]	8,80%	95%	5,19

Fonte: elaborada pelo autor (2021)

Figura 82 - Caso de uso - Informações da prova

ANTECEDENTE(S)	CONSEQUENTE	FREQUÊNCIA	CONFIANÇA	ALAVANCAGEM
[Nota da redação: entre 750 e 1000] [Língua estrangeira: inglês]	Média total das notas entre 400 e 600	7,20%	91,90%	2,26
[Nota da redação: entre 250 e 500] [Língua estrangeira: inglês]	Redação normal(não zerada)	7,50%	100%	1,35
[Nota da redação: entre 250 e 500] [Língua estrangeira: espanhol]	Redação normal(não zerada)	11,90%	100%	1,35
[Falta na redação] [Língua estrangeira: espanhol]	Falta em uma das áreas de conhecimento	13,40%	99,90%	3,65
[Nota da redação: entre 500 e 750] [Média total das notas: entre 400 e 600]	Redação normal(não zerada)	21,20%	99,9	1,35
[Nota da redação: entre 750 e 1000]	Média total das notas entre 400 e 600	10,40%	87,90%	2,5
[Nota da redação: entre 750 e 1000] [Língua estrangeira: inglês]	Redação normal(não zerada)	7,80%	100%	1,35

Fonte: elaborada pelo autor (2021)

Figura 83 - Caso de uso - Localidade

ANTECEDENTE(S)	CONSEQUENTE	FREQUÊNCIA	CONFIANÇA	ALAVANCAGEM
[Nota da redação: entre 750 e 1000] [UF da prova: SP]	Média total das notas entre 400 e 600	1,80%	93,20%	2,65
[Nota da redação: entre 750 e 1000] [UF da prova: MG]	Média total das notas entre 400 e 600	1,50%	91,40%	2,6
[Nota da redação: entre 250 e 500] [UF da prova: PA]	Média total das notas entre 200 e 400	1%	83,80%	2,2
[Nota da redação: entre 250 e 500] [UF da prova: BA]	Média total das notas entre 200 e 400	1,30%	81,40%	2,17

Fonte: elaborada pelo autor (2021)

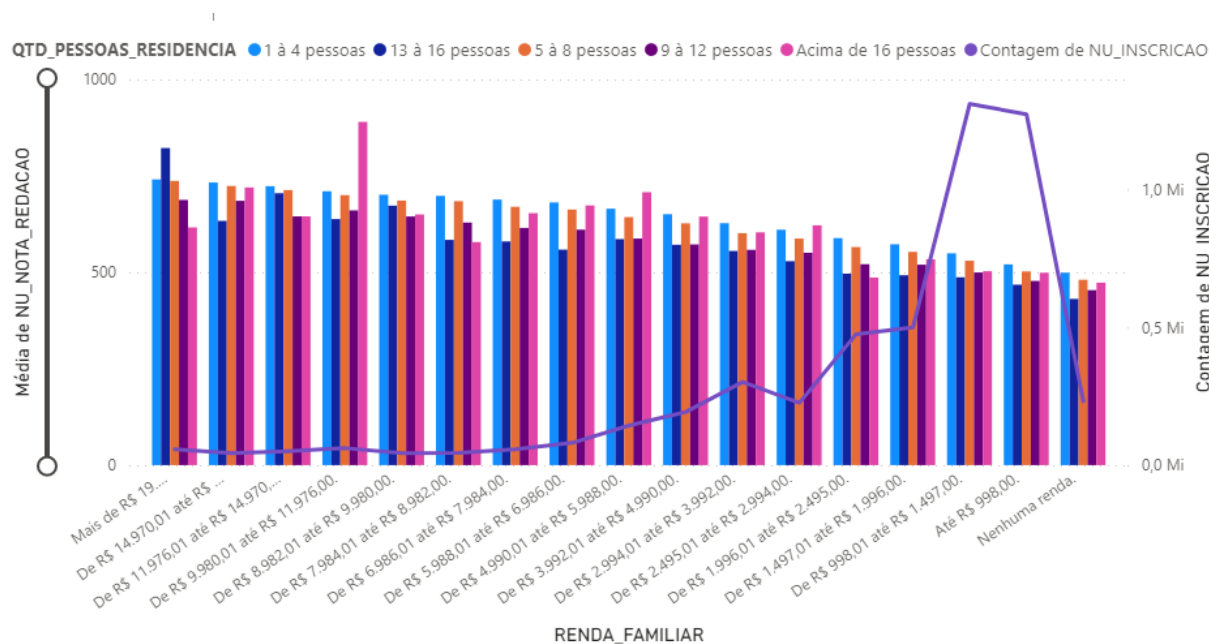
4.3 INTERPRETAÇÃO DOS RESULTADOS PELOS ESPECIALISTAS EM ASSISTÊNCIA SOCIAL E EDUCAÇÃO.

Baseado nas opiniões dos especialistas vimos que os principais fatores que podem influenciar positivamente ou negativamente no desempenho de uma participante está na renda familiar, tipo de escola e quantidade pessoas, essas informações foram possíveis confirmar

através dos dados correlacionados. E a partir dessa informação podemos atestar que quanto maior a renda maior, melhor é o desempenho na redação, a quantidade de pessoas impacta no despenho da nota na maioria dos casos, mas com a renda maior, esse quesito evolui também. Podemos destacar também que a maior parte dos inscritos se concentra nas primeiras faixas de rendas familiares.

Na figura 84 podemos ver melhor essa exemplificação.

Figura 84 -Gráfico de pareto exibindo uma comparação entre as rendas familiares e as aplicando as médias da nota redação a quantidade de pessoas por residência.

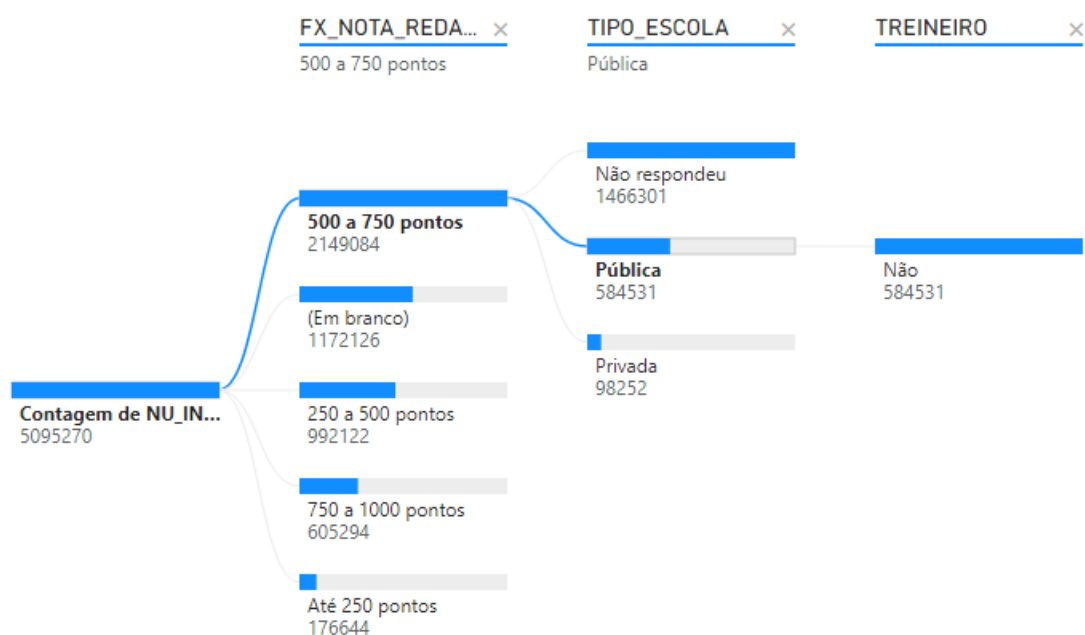


Fonte: Elaboração do autor (2021)

Podemos destacar a situação do ensino público e privado que foi constatado pelo especialista em entrevista¹ que relatou a diferença do currículo escolar que não está no mesmo nível, porém o que o exame do ENEM exige a maioria dos colégios de ensino médio público não cobrem, logo para que um participante possa estar preparado é necessário estudar em um lugar mais qualificado ou fazer um pré-vestibular o que vai demandar tempo e em muitas ocasiões pode demandar custo. Na figura 85 podemos verificar que a maioria dos praticantes que tiram uma nota na faixa de 500 a 750 pontos na redação, que é a mais alta, são a maioria, entretanto na figura 86 podemos verificar que a média dos participantes de escola privada tem a média bastante superior quando comparados.

¹ Entrevista realizada com especialista Damião Santos formado na Área Social

Figura 85 - Gráfico de árvore hierárquica que demonstra o caminho percorrido e aplicando a quantidade de participantes que está correlacionado a cada atributo



Fonte: Elaboração do autor (2021)

Figura 86 - Gráfico de árvore hierárquica que demonstra o caminho percorrido e aplica a média da redação para cada atributo correlacionado.



Fonte: Elaboração do autor (2021)

Com essas informações podemos aferir que um grande desequilíbrio entre o ensino público e privado a respeito da preparação para o ENEM.

4.4 COMPARAÇÃO ENTRE AS ANÁLISES E O SEGUNDO QUESTIONÁRIO DOS ESPECIALISTAS.

No segundo questionário com os especialistas da área de educação e assistência social, foram apresentados em linhas gerais os itens da base a fim de levantar quais deles teriam mais relevância para associar com as medidas de desempenho do inscrito. Foram obtidas quatro respostas e as tendências foram analisadas, e nesta seção é realizada a comparação entre o que os profissionais definiram como itens apropriados para associar ao desempenho e quais os itens foram de fato mais determinantes nas associações.

A maioria dos especialistas classificou a prioridade das informações pessoais do participante em relação ao seu impacto no desempenho, da maior para menor relevância, em Localidade de residência, localidade do nascimento, idade, cor(raça), gênero e nacionalidade. Em se tratando de grupos de antecedentes para associar as métricas de desempenho, a maior parte apontou o grupo de local de nascimento e cor.

Sobre as informações de conclusão de ensino médio, os grupos de associação não tiveram tendência e foram bem diversificados.

A classificação geral da relevância das informações de ensino médio, da maior relevância para a menor, é dada por localidade da escola, tipo da localização, dependência administrativa e situação de funcionamento. Todos eles apontaram como grupo antecedente da associação o conjunto composto pelo par situação de operação e dependência administrativa da escola.

Sobre os questionários socioeconômicos, a ordem de classificação dos itens de situação familiar em relação a seu impacto em desempenho é, do maior ao menor, renda familiar, formação do pai, profissão do pai, formação da mãe, quantidade de pessoas em casa. A maioria determinou os grupos de associação que compreendem os conjuntos de escolaridade e formação do pai, escolaridade e formação da mãe e a renda familiar com a quantidade de pessoas.

Sobre a situação doméstica, a ordem de classificação da maioria foi indo daquela mais para a menos relevante, empregada doméstica, quarto, carro, banheiro e moto. Os grupos de associação mais frequentes foram empregada e quarto, carro e moto, moto e carro. Nos eletrodomésticos, os 3 itens mais relevantes classificados, da menor para maior relevância,

foram micro-ondas, freezer e máquina de lavar. Os grupos de antecedentes mais comuns foram máquina de lavar e de secar, geladeira e freezer.

Nos grupos de aparelhos de multimídia e telecomunicações, a ordem de relevância decrescente dos itens que impactem no desempenho são computador, acesso a internet e celular. Sobre os grupos de antecedência, todos eles apontaram o conjunto de internet e computador, TV a cores e por assinatura, DVD e TV a cores.

Sobre todos os itens, foi perguntado quais deles, que correspondem aos casos de uso da mineração, são os mais relevantes para focar as análises. A maioria respondeu os grupos de renda familiar, quantidade de pessoas na casa, itens residenciais e eletrodomésticos.

Na mineração e através da análise de dados, foi possível contemplar as regras de associação e testar hipóteses com base nelas em concordância com o que a maioria dos especialistas julgam relevante. A discrepância está nos eletrodomésticos, que não foram identificados como significantes do impacto.

5. CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS

O objetivo deste trabalho foi realizar um estudo sobre o impacto dos atributos socioeconômicos ao rendimento do ao rendimento dos participantes do ENEM, utilizando ferramentas de mineração de dados e de inteligência de negócio.

O primeiro passo do trabalho foi coletar uma base de dados recente e pública que contivesse dados sobre educação de base para ser analisada sob a ótica da inteligência de negócios e do fluxo KDD, para apoiar tomada de decisões orientada a dados. A solução encontrada foi utilizar a base de microdados do Enem 2019, disponível no portal do Inep, a mais recente disponibilizada. Esta base foi baixada, carregada para um repositório em nuvem do tipo “lago de dados”, que permite democratizar os dados para suportar diversos formatos e conectar-se a diversas fontes, no sentido de um local comum de onde todos os dados vem e para todos eles vão.

Os dados brutos foram movimentados deste armazenamento em nuvem para bancos de dados em nuvem da Azure, onde se realizou a homologação do modelo para desenvolvimento e posteriormente para um banco de dados local Microsoft SQL Server, onde se desenvolveu o modelo de dados de produção, arquitetado no modelo “mercado de dados relacional”, no qual foi extraído para documentos carregados novamente para área de preparação do repositório lago de dados para serem usados na análise de dados, e no banco de dados modelado, foi aplicado

pré-processamento e transformação para extrair o conjunto de dados para a mineração, que foi armazenado localmente para aplicar o algoritmo de regras de associação apriori.

A mineração de e a análise de dados do tipo ad-hoc foram aplicados em concomitância nos modelos e no conjunto de dados preparados através dos fluxos de ETL e ELT, respectivamente. Através da descoberta de regras de associação, foi possível varrer os conjuntos e descobrir hipóteses reveladas a partir de limiares pré-definidas em parâmetros de probabilidade e estatística. Na análise de dados, foi realizado o levantamento geral das estatísticas e também o teste das hipóteses encontradas com a mineração, além das que foram levantadas em um questionário e entrevistas com especialistas para levantar requisitos para essas avaliações e aferir quais dados são mais relevantes para analisar impacto no desempenho do exame.

Unificando as informações da mineração e opinião dos especialistas foi possível identificar os pontos-chaves que exibem as diferenças sociais que podem afetar o rendimento dos participantes do ENEM. Com isso foi possível desmistificar alguns pontos que poderiam ser interpretados como ruim para o desempenho dos participantes. Essas relações de informações foram verificadas analisando a média de desempenho das notas de redação e prova objetiva e a quantidade de inscrito por atribuição dos dados.

Segundo os especialistas o principal ponto é a renda familiar, esse item impacta em todos os aspectos socioeconômicos e inevitavelmente afeta o desempenho do participante, seja para um bom desempenho ou para um mau desempenho. Durante a aplicação da análise descritiva através da mineração e do teste de hipóteses nos pós-processamento dos dados minerados, as variáveis de maior impacto no desempenho do participante são a escolaridade dos pais e o tipo de escola em que se cursou o ensino médio. Bem como apontado pelos profissionais especializados, os fatores preponderantes no rendimento nas provas e redação são, majoritariamente, a renda mensal familiar e a ocupação dos pais.

Constatou-se que o, assim como indicado pelos profissionais, o número de pessoas na casa, se for baixo, tende a se relacionar com bons desempenhos nas notas de redação e média das provas objetivas, enquanto os outros parâmetros como dados de renda e ocupação foram controlados. Quando as variáveis como nível de renda, grupo ocupacional e nível educacional dos pais assumem valores de altos patamares, que são menos frequentes, a variável da quantidade de pessoas na casa tende a ter impacto ínfimo.

Conclui-se a partir das análises que, as disparidades socioeconômicas impactam de forma determinante nos graus de desempenho dos alunos, onde nota-se que as desigualdades

de renda, de ocupação, de acesso a escolas de nível educacional melhor e que residem em locais onde a renda regional é mais baixa tendem a se sair pior no exame, de uma forma geral. Também é possível salientar que características pessoais específicas como sexo, estado civil e tipo racial tendem a impactar na performance do participante, enquanto relacionadas segundo determinados perfis. Além disto, a idade dos participantes também é um fator que tem relevância nesta análise.

Os conhecimentos gerados com base na mineração de dados evidenciam a ocorrência destes fatos sociais que indicam desigualdades, ao passo que com eles em mãos e a partir dos testes das hipóteses geradas deles pelas regras de associação com desempenho no exame e com o apoio interpretativo especializado de profissionais da assistência social e educação, é possível produzir uma sabedoria através de cada conhecimento coletado com base na estatística descritiva, assim aplicando uma análise diagnóstica dos fatores de impacto nas métricas das provas realizadas e as explicações válidas por trás deles, de modo a perfilar as medidas de acordo com as informações dos participantes.

Com este trabalho, foi possível diagnosticar indicativos de desigualdade social nos participantes do ENEM no ano de 2019 e como estes influenciam em seus resultados no exame, com o apoio teórico especializado para avaliar as conclusões e na aplicação da avaliação interpretativa final. Assim, é possível provar a estes profissionais da área social, pedagógica e educacional um ferramental bruto, consistente e evolutivo para consulta e estabelecimento de análises técnicas e pesquisas aplicadas, para levantar correlações e inferir causalidades a partir dos dados provisionados.

Desta forma, conclui-se que o projeto de inteligência de negócios modernos pautado em Big Data é uma ferramenta válida, poderosa e altamente escalável, adaptável e customizável, que permite empregar um ciclo de vida de análise de dados massivos muito amplo e com emprego de diversas tecnologias robustas para viabilizar casos de uso mais complexos. Tendo isto posto, trata-se de uma metodologia crucial e de valor agregado considerável para desenvolver sistemas e ferramentas para apoio a tomada de decisão baseadas em dados a partir da análise de mega dados.

5.1 TRABALHOS FUTUROS

Dada a vastidão metodológica e teórica presente neste trabalho, há diversas oportunidades de outros projetos futuros baseados neste modelo ferramental escolhido, com aplicações processuais e operacionais para diversos outros assuntos do mesmo cunho, mas

também para tópicos de outras áreas dos estudos sociais. Aqui apresentamos algumas possibilidades principais.

- **Estudo da educação inclusiva e acessibilidade**

O mesmo fluxo de processos pode ser aplicado para avaliar o exame nacional do ensino médio sob a ótica da inclusão social e acessibilidade, analisando apenas ou com foco maior os dados referentes as condições de atendimento especial ou específico na prova. Há uma vasta quantidade de dados e informações úteis possíveis de se extrair deles disponíveis nas bases de microdados.

- **Estudo do censo escolar de escolas públicas**

Pode-se utilizar as ferramentas e processos aplicados neste trabalho para o estudo dos censos escolares, especificamente realizados em escolas públicas, para a extração e entendimento dos dados e buscar compreendê-los através da abordagem de descoberta de conhecimento em bases de dados.

- **Estruturação e criação de um projeto de business intelligence fim a fim para uma rede de escolas de educação básica ou técnica**

O modelo arquitetural completo de inteligência de negócios com a metodologia de Descoberta de Conhecimento em Bases de Dados é altamente recomendável para ser aplicado em um ambiente de estudos educacionais em uma rede escolar de educação básica regular ou técnica e até mesmo pode contemplar a análise de escolas técnicas da modalidade subsequente. Assim, há a possibilidade de se detectar, gerenciar e garantir a governança dos dados na instituição, bem como a democratização dos dados para os diversos profissionais da área analítica, que os utilizarão para diversos fins

REFERÊNCIAS BIBLIOGRÁFICAS

JOEDAVIES-MSFT. **Assinaturas, licenças, contas e locatários para ofertas de nuvem da Microsoft - Microsoft 365 Enterprise**. Disponível em: <<https://docs.microsoft.com/pt-br/microsoft-365/enterprise/subscriptions-licenses-accounts-and-tenants-for-microsoft-cloud-offerings?view=o365-worldwide>>. Acesso em: 11 maio. 2021.

Azure for College Students — Detalhes da oferta | Microsoft Azure. Disponível em:

<<https://azure.microsoft.com/pt-br/offers/ms-azr-0170p/>>. Acesso em: 11 maio. 2021.

Preços – Data Lake Store Gen1 | Microsoft Azure. Disponível em:

<<https://azure.microsoft.com/pt-br/pricing/details/data-lake-storage-gen1/>>. Acesso em: 11 maio. 2021.

Preço – Banco de Dados SQL Individual do Azure | Microsoft Azure. Disponível em:

<<https://azure.microsoft.com/pt-br/pricing/details/azure-sql-database/single/>>. Acesso em: 11 maio. 2021.

Preços do Analysis Services | Microsoft Azure. Disponível em:

<<https://azure.microsoft.com/pt-br/pricing/details/analysis-services/>>. Acesso em: 11 maio. 2021.

MOREIRA, F. P. A INFLUÊNCIA DA SEPARAÇÃO DOS PAIS NO DESEMPENHO ESCOLAR DE ALUNOS DAS SÉRIES/ANOS INICIAIS NA VISÃO DO

PROFESSOR. Trabalho de Conclusão de Curso—Universidade do Extremo Sul Catarinense:

[s.n.].

INEP. **Dicionário Microdados ENEM 2019**Dicionário Microdados ENEM 2019. [s.l.]

INEP, 20 fev. 2020.

BRASIL. **Lei Geral de Proteção de Dados Pessoais (LGPD)**Planalto.gov.br, 14 ago. 2018.

Disponível em: <https://www.planalto.gov.br/ccivil_03/_ato2015-

2018/2018/lei/L13709.htm>. Acesso em: 24 maio. 2021

O que é o INEP e qual a sua responsabilidade? Tudo sobre Vestibulares e ENEM!

Disponível em: <<https://dicas.vestibulares.com.br/inep/>>. Acesso em: 26 abril 2021.

Descubra o que é o Inep e o Encceja e para o que eles servem. Guia da Carreira.

Disponível em: <<https://www.guiadacarreira.com.br/educacao/inep-encceja/>>. Acesso em: 4 May 2021.

Entenda rapidamente o que é o ENEM. Guia da Carreira. Disponível em:

<<https://www.guiadacarreira.com.br/educacao/o-que-e-enem/>>. Acesso em: 4 May 2021.

EDUCA MAIS BRASIL. **Educa Mais Brasil - Bolsas de Estudo de até 70% para**

Faculdades – Graduação e Pós-graduação. Educa Mais Brasil. Disponível em:

<<https://www.educamaisbrasil.com.br/programas-do-governo/enem/o-que-e->>. Acesso em: 4 May 2021.

O que é o Enem e como funciona? - Tudo Sobre Enem. Tudo Sobre Enem. Disponível em: <<https://descomplica.com.br/tudo-sobre-enem/enem/o-que-e-o-enem/>>. Acesso em: 4 May 2021.

TANCREDI, Silvia. **Atendimento especializado no Enem - como funciona.** Super Vestibular. Disponível em: <<https://vestibular.mundoeducacao.uol.com.br/enem/atendimento-especial-no-enem-como-funciona.htm>>. Acesso em: 4 May 2021.

PRASABER. **Como pedir a isenção do Enem 2020?** Pravalor | Soluções em Crédito Universitário. Disponível em: <<https://www.pravalor.com.br/como-pedir-a-isencao-do-enem-2020/#:~:text=Os%20inscritos%20no%20Cad%C3%9Anico%20s%C3%A3o,seja%20feita%20na%20Receita%20Federal.>>. Acesso em: 4 May 2021.

BLOG DO EAD UCS. **Big Data: O Que é, Para Que Serve, Como Aplicar e Exemplos.** Disponível em: <<https://ead.ucs.br/blog/big-data>>. Acesso em: 13 maio. 2021.

NASCIMENTO, R. **O que é Big Data e para que ele serve? | Marketing por Dados.** Disponível em: <<http://marketingpordados.com/analise-de-dados/o-que-e-big-data-%F0%9F%A4%96/>>. Acesso em: 13 maio. 2021.

RUSSOM, P. **BIG DATA ANALYTICS Vivomente.** [s.l.] Vivomente, 2011. Disponível em: <<https://vivomente.com/wp-content/uploads/2016/04/big-data-analytics-white-paper.pdf>>. Acesso em: 13 maio. 2021.

GOMES, L. **Big Data Analytics: você sabe o que é e como funciona? | Blog Voitto.** Disponível em: <<https://www.voitto.com.br/blog/artigo/big-data-analytics>>. Acesso em: 13 maio. 2021.

Data Scientist, Data Analyst e Data Engineer: Tem Diferença? Disponível em: <<https://bigdatacorp.com.br/data-scientist-data-analyst-e-data-engineer-tem-diferenca/>>. Acesso em: 14 maio. 2021.

GARCIA, M. **Data Engineer ou Engenheiro de Dados – Conheça mais sobre.** Disponível em: <<https://www.cetax.com.br/blog/data-engineer-ou-engenheiro-de-dados/>>. Acesso em: 14 maio. 2021.

Big Data Engineer: Role, Responsibilities, and Job Description. AltexSoft. Disponível em: <<https://www.altexsoft.com/blog/big-data-engineer/>>. Acesso em: 12 May 2021.

What is Data Engineer: Role Description, Responsibilities, Skills, and Background. AltexSoft. Disponível em: <<https://www.altexsoft.com/blog/what-is-data-engineer-role-skills/>>. Acesso em: 12 May 2021.

What is Data Pipeline: Components, Types, and Use Cases. Disponível em:

<<https://www.altexsoft.com/blog/data-pipeline-components-and-types/>>. Acesso em: 17 maio. 2021.

SHIFF LAURA. **Real Time vs Batch Processing vs Stream Processing.** Disponível em:

<<https://www.bmc.com/blogs/batch-processing-stream-processing-real-time/>>. Acesso em: 16 maio. 2021.

KELLISON FERREIRA. **Análise de dados: o que é, como fazer e dicas básicas.** Rock

Content - BR. Disponível em: <<https://rockcontent.com/br/blog/analise-de-dados/>>. Acesso em: 4 May 2021.

PORTO, F.; ZIVIANI, A. **Ciência de Dados.** [s.l.] , [s.d.]. Disponível em:

<<https://www.lncc.br/~ziviani/papers/III-Desafios-SBC2014-CiD.pdf>>.

Big Data: você conhece os 4 tipos de análise de dados? – Blog Academia IN. Disponível

em: <<https://blog.academiain1.com.br/big-data-voce-conhece-os-4-tipos-de-analise-de-dados/>>. Acesso em: 22 maio. 2021.

REIS, F. **O que é uma consulta Ad Hoc em bancos de dados - Bóson Treinamentos em Ciência e Tecnologia.** Disponível em: <<http://www.bosontreinamentos.com.br/bancos-de-dados/o-que-e-uma-consulta-ad-hoc-em-bancos-de-dados/>>. Acesso em: 13 maio. 2021.

PORTO, Fábio ; ZIVIANI, Artur. **Ciência de Dados.** [s.l.]: , [s.d.]. Disponível em:

<<https://www.lncc.br/~ziviani/papers/III-Desafios-SBC2014-CiD.pdf>>.

Introdução à ciencia de dados, Fernando Amaral, mineração de dados e Big data, Pag 3, 2016,

Data de acesso 29/04/2021 COELHO, Lucas. **Ciência de Dados: O que é, conceito e**

definição | Blog Cetax. Data Analytics, Big Data, Data Science - Blog Cetax. Disponível em:

<<https://www.cetax.com.br/blog/data-science-ou-ciencia-de-dados/>>. Acesso em: 10 May 2021.

Vista do Big Data, Data Science e seus contributos para o avanço no uso da Open Source Intelligence | Sistemas & Gestão. Revistasg.uff.br. Disponível em:

<<https://revistasg.uff.br/sg/article/view/1026/547>>. Acesso em: 10 May 2021.

Data Science para Negócios. Google Books. Disponível em:

<[https://books.google.com.br/books?hl=pt-](https://books.google.com.br/books?hl=pt-BR&lr=lang_pt&id=1rZwDwAAQBAJ&oi=fnd&pg=PT8&dq=data+science&ots=MxLj6Wy4Ro&sig=k25IrsokJ_8MZNhIID9-HyT4W5k#v=onepage&q&f=false)

[BR&lr=lang_pt&id=1rZwDwAAQBAJ&oi=fnd&pg=PT8&dq=data+science&ots=MxLj6Wy4Ro&sig=k25IrsokJ_8MZNhIID9-HyT4W5k#v=onepage&q&f=false](https://books.google.com.br/books?hl=pt-BR&lr=lang_pt&id=1rZwDwAAQBAJ&oi=fnd&pg=PT8&dq=data+science&ots=MxLj6Wy4Ro&sig=k25IrsokJ_8MZNhIID9-HyT4W5k#v=onepage&q&f=false)>. Acesso em: 10 May 2021.

What is business intelligence? Your guide to BI and why it matters. Tableau. Disponível em: <<https://www.tableau.com/learn/articles/business-intelligence>>. Acesso em: 10 May 2021.

Desmistificando o BI Conceitos, estruturas e principais ferramentas – DBCCOMPANY. Disponível em: <<https://www.dbccompany.com.br/wp-content/uploads/2019/07/Desmistificando-bi.pdf>> Acesso em: 12 May 2021.

UMA ARQUITETURA PARA BUSINESS INTELLIGENCE BASEADA EM TECNOLOGIAS SEMÂNTICAS PARA SUPORTE A APLICAÇÕES ANALÍTICAS.

[s.l.] , 2006. Disponível em:

<<https://www.cin.ufpe.br/~sfa/Uma%20Arquitetura%20Para%20Business%20Intelligence%20Baseada%20Em%20Tecnologias%20Sem%20E2nticas%20Para%20Suporte%20A%20Aplica%20E7%F5es%20Anal%20EDticas.pdf>>. Acesso em: 21 maio. 2021.

SANTOS. Business Intelligence 2.0: Diferença entre o BI tradicional e o BI 2.0.

DevMedia. Disponível em: <<https://www.devmedia.com.br/business-intelligence-2-0-conceitos-componentes-e-arquitetura/28899>>. Acesso em: 13 May 2021.

Leandro Guimarães. Disponível em: <<https://www.knowsolution.com.br/arquitetura-bi-como-funciona-como-trabalhar-conceito/#:~:text=A%20arquitetura%20de%20Business%20Intelligence,construir%20os%20sistemas%20de%20BI.>>. Acesso em: 13 maio. 2021.

Arquitetura do BI | BI NA PRÁTICA. Disponível em:

<<https://www.binapratica.com.br/arquitetura-bi>>. Acesso em: 13 maio. 2021.

Data-Driven Design

BRACARENSE, Leonardo. Data-driven Design: entenda o que é e como ele pode te ajudar | Marketing por Dados. Marketing por Dados. Disponível em:

<<http://marketingpordados.com/analise-de-dados/data-driven-design-entenda-o-que-e-e-como-ele-pode-te-ajudar/>>. Acesso em: 13 May 2021.

Mineração de dados

O que é mineração de dados? Sas.com. Disponível em:

<https://www.sas.com/pt_br/insights/analytics/mineracao-de-dados.html>. Acesso em: 10 May 2021.

RIBEIRO, Jefferson. O que é Data Mining: Conceitos e Técnicas sobre Data Mining.

DevMedia. Disponível em: <<https://www.devmedia.com.br/conceitos-e-tecnicas-sobre-data-mining/19342>>. Acesso em: 10 May 2021.

AMORIM, Thiago. **Conceitos, técnicas, ferramentas e aplicações de Mineração de Dados para gerar conhecimento a partir de bases de dados**. [s.l.]: , 2006. Disponível em:

<<https://www.cin.ufpe.br/~tg/2006-2/tmas.pdf>>. Acesso em: 10 May 2021.

Data Science x Business Intelligence

Data Science vs. Business Intelligence: What's the Difference? online.sju. Disponível em:

<[https://online.sju.edu/graduate/masters-business-intelligence/resources/articles/data-science-or-business-](https://online.sju.edu/graduate/masters-business-intelligence/resources/articles/data-science-or-business-intelligence#:~:text=In%20a%20nutshell%2C%20BI%20analysts,make%20predictions%20for%20the%20future.&text=BI%20analysts%2C%20on%20the%20other%20hand%2C%20interpret%20past%20trends.>)

[intelligence#:~:text=In%20a%20nutshell%2C%20BI%20analysts,make%20predictions%20for%20the%20future.&text=BI%20analysts%2C%20on%20the%20other%20hand%2C%20interpret%20past%20trends.>](https://online.sju.edu/graduate/masters-business-intelligence/resources/articles/data-science-or-business-intelligence#:~:text=In%20a%20nutshell%2C%20BI%20analysts,make%20predictions%20for%20the%20future.&text=BI%20analysts%2C%20on%20the%20other%20hand%2C%20interpret%20past%20trends.>). Acesso em: 24 Apr 2021.

Business Intelligence vs. Data Mining: A Comparison - Talend. Talend Real-Time Open Source Data Integration Software. Disponível em:

<<https://www.talend.com/resources/business-intelligence-data-mining/>>. Acesso em: 10 May 2021.

DIKW

FIGUEROA, A. **Data Demystified — DIKW model - Towards Data Science**. Disponível em: <<https://towardsdatascience.com/rootstrap-dikw-model-32cef9ae6dfb>>. Acesso em: 16 maio. 2021.

PAULO MARTINS. **A Pirâmide do Conhecimento e Hierarquia DIKW | TargetData**.

Disponível em: <<https://www.targetdata.com.br/higienizacao-e-enriquecimento-de-dados/piramide-do-conhecimento>>. Acesso em: 16 maio. 2021.

MATHEUS BATISTA FURLAN ALGORITMOS E TÉCNICAS PARA MINERAÇÃO DE DADOS. [s.l.] , 2018. Disponível em:

<<https://cepein.femanet.com.br/BDigital/arqTccs/1511420203.pdf>>.

Técnicas de Mineração de dados

MINEWISKAN. **Algoritmos de mineração de dados (Analysis Services-Mineração de dados)**. Disponível em: <<https://docs.microsoft.com/pt-br/analysis-services/data-mining/data-mining-algorithms-analysis-services-data-mining?view=asallproducts-allversions>>. Acesso em: 20 abril. 2021.

DE AMO, S. **Técnicas de Mineração de Dados**. [s.l.] , [s.d.]. Disponível em:

<<http://files.sistemas2012.webnode.com.br/2000000095-bf367bfb43/Tecnicas%20de%20Minera%C3%A7%C3%A3o%20de%20Dados.pdf>>. Acesso em: 20 abril. 2021.

DUNHAM, M.; LE GRUENWALD, Y.; HOSSAIN, Z. **A SURVEY OF ASSOCIATION RULES**. [s.l.] , [s.d.]. Disponível em: <<http://www2.cs.uh.edu/~ceick/6340/grue-assoc.pdf>>.

Acesso em: 17 maio. 2021.

BOUBAKER, B.; OURIDA, S.; WAFI, T. Formal Concept Analysis Based Association Rules Extraction. **IJCSI International Journal of Computer Science Issues**, v. 8, n. 2, p. 1694-0814, 2011.

SHWETA, Ms ; GARG, Kanwal. Mining Efficient Association Rules Through Apriori Algorithm Using Attributes and Comparative Analysis of Various Association Rule Algorithms. v. 3, n. 6, 2013. Disponível em:

<http://www.ting tongb2b.com/aims/pdf/Apriori_Main.pdf>.

PÔSSAS, B. et al. Using quantitative information for efficient association rule generation.

Journal of the Brazilian Computer Society, v. 7, n. 1, p. 35–42, 2000.

EDUARDO CORRÊA GONÇALVES. **Data Mining de Regras de Associação**. Disponível em: <<https://www.devmedia.com.br/data-mining-de-regras-de-associao/6941>>. Acesso em: 17 maio. 2021.

CETAX. **Big Data: O que é, conceito e definição | Blog Cetax**. Disponível em:

<<https://www.cetax.com.br/blog/big-data/>>. Acesso em: 13 maio. 2021.

Big Data Analytics

LEONNARDO GOMES. **Big Data Analytics: você sabe o que é e como funciona? | Blog Voitto**. Disponível em: <<https://www.voitto.com.br/blog/artigo/big-data-analytics>>. Acesso em: 13 maio. 2021.

Data de acesso 24/04/2021

CETAX. **ETL - Extract Transform Load : O que é, conceitos e definição | Blog Cetax**.

Data Analytics, Big Data, Data Science - Blog Cetax. Disponível em:

<<https://www.cetax.com.br/blog/etl-extract-transform-load/>>. Acesso em: 11 May 2021.

ISABELA BLASI. **ETL X ELT : qual a diferença?** Disponível em:

<[https://blog.indicium.tech/etl-vs-elt-](https://blog.indicium.tech/etl-vs-elt-diferencas/#:~:text=Apesar%20disso%2C%20com%20o%20surgimento,nas%20opera%C3%A7%C3%B5es%20modernas%20de%20dados.)

[diferencas/#:~:text=Apesar%20disso%2C%20com%20o%20surgimento,nas%20opera%C3%A7%C3%B5es%20modernas%20de%20dados.](https://blog.indicium.tech/etl-vs-elt-diferencas/#:~:text=Apesar%20disso%2C%20com%20o%20surgimento,nas%20opera%C3%A7%C3%B5es%20modernas%20de%20dados.)>. Acesso em: 15 maio. 2021.

MIRANDA, W. **Modelagem de Dados: O que é e para que serve para um DBA**.

Disponível em: <<https://www.portalgsti.com.br/2017/02/modelagem-de-dados-o-que-e-e-para-que-serve-para-um-dba.html>>. Acesso em: 16 maio. 2021.

RODRIGUES, Joel. **MER e DER: Modelagem de Bancos de Dados**. DevMedia. Disponível em: <<https://www.devmedia.com.br/modelo-entidade-relacionamento-mer-e-diagrama-entidade-relacionamento-der/14332>>. Acesso em: 5 May 2021.

SAMUEL, N. **Star and Snowflake Schema: A Comprehensive Analysis**. Disponível em: <<https://hevodata.com/learn/star-and-snowflake-schema-analysis/>>. Acesso em: 15 maio. 2021.

Star Schema in Data Warehouse modeling - GeeksforGeeks. Disponível em: <<https://www.geeksforgeeks.org/star-schema-in-data-warehouse-modeling/>>. Acesso em: 16 maio. 2021.

DATA WAREHOUSE Análise de dados (Big Data). [s.l.]: , [s.d.]. Disponível em: <<https://docente.ifrn.edu.br/josecunha/disciplinas/adbd/pdfs/data-warehouse>>. Acesso em: 5 May 2021.

MOURA, L. **Data Warehouse ou Data Mart: Por eu devo começar?** Disponível em: <<https://www.devmedia.com.br/data-warehouse-ou-data-mart-por-onde-comecar/6996>>. Acesso em: 15 maio. 2021.

DATA WAREHOUSING, DATA MINING, OLAP AND OLTP TECHNOLOGIES ARE ESSENTIAL ELEMENTS TO SUPPORT DECISION-MAKING PROCESS IN INDUSTRIES. citeseerx.ist.psu.edu. Disponível em: <<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.301.5615&rep=rep1&type=pdf>>.

JATIN RAISINGHANI. **Data Lake vs Data Warehouse vs Data Mart**. Disponível em: <<https://www.holistics.io/blog/data-lake-vs-data-warehouse-vs-data-mart/>>. Acesso em: 13 maio. 2021.

DAVID MATOS. **Data Lake, a fonte do Big Data — Ciência e Dados**. Disponível em: <<https://www.cienciaedados.com/data-lake-a-fonte-do-big-data/>>. Acesso em: 15 maio. 2021.

Datawarehouse moderno e tradicional

DESENVOLVIDO POR ROCK CONTENT. **Modern Data Warehouse: entenda tudo sobre esse conceito - Arbit**. Disponível em: <<https://blog.arbit.com.br/modern-data-warehouse-entenda-conceito/>>. Acesso em: 15 maio. 2021.

Modelos de processamento de dados, OLTP, OLAP

OLAP vs. OLTP: What's the Difference? Ibm.com. Disponível em: <<https://www.ibm.com/cloud/blog/olap-vs-oltp>>. Acesso em: 12 May 2021.

CETAX. **O que é OLAP - Online Analytical Processing ? | Blog Cetax.** Data Analytics, Big Data, Data Science - Blog Cetax. Disponível em: <<https://www.cetax.com.br/blog/o-que-e-olap/>>. Acesso em: 12 May 2021.

WALDES OLIVEIRA. **Diferenças entre OLTP e OLAP.** TechTem. Disponível em: <<https://www.techtem.com.br/diferencas-entre-oltp-e-olap/>>. Acesso em: 12 May 2021.

MOURA, Luis. **O que é OLAP? Conceitos Básicos Sobre OLAP.** DevMedia. Disponível em: <<https://www.devmedia.com.br/conceitos-basicos-sobre-olap/12523>>. Acesso em: 12 May 2021.

URIONA, M. **Business Intelligence.** Disponível em: <<https://pt.slideshare.net/MauricioUrionaMaldon/business-intelligence-75401895>>. Acesso em: 13 maio. 2021.

MINEWISKAN. **Comparing Analysis Services tabular and multidimensional models.** Microsoft.com. Disponível em: <<https://docs.microsoft.com/en-us/analysis-services/comparing-tabular-and-multidimensional-solutions-ssas?view=asallproducts-allversions>>. Acesso em: 12 May 2021.

CALBIMONTE, Daniel. **Tabular vs Multidimensional models for SSAS.** Mssqltips.com. Disponível em: <<https://www.mssqltips.com/sqlservertip/4154/tabular-vs-multidimensional-models-for-sql-server-analysis-services/>>. Acesso em: 12 May 2021.

Vista do Cloud services e o padrão PREMIS. Disponível em: <<https://periodicos.sbu.unicamp.br/ojs/index.php/rdbci/article/view/8661384/25825>>. Acesso em: 22 maio. 2021.

What is a Cloud Service? – Cloud Service Definition - Citrix. Disponível em: <<https://www.citrix.com/solutions/digital-workspace/what-is-a-cloud-service.html>>. Acesso em: 22 maio. 2021.

Data de acesso 29/04/2021

https://ava.unicarioca.edu.br/graduacao/pluginfile.php/969791/mod_resource/content/0/Tema%206%20-%20Resumo.pdf

CHAPPELL, David. **INTRODUCING THE AZURE SERVICES PLATFORM AN EARLY LOOK AT WINDOWS AZURE, .NET SERVICES, SQL SERVICES, AND LIVE SERVICES SPONSORED BY MICROSOFT CORPORATION 2 CONTENTS.**

[s.l.]: , 2009. Disponível em:

<http://www.davidchappell.com/Azure_Services_Platform_v1.1--Chappell.pdf>. Acesso em: 11 May 2021.

JONBURCHEL. **Documentação do Azure Data Factory - Azure Data Factory.**

Microsoft.com. Disponível em: <<https://docs.microsoft.com/pt-br/azure/data-factory/>>.

Acesso em: 12 May 2021.

Azure Data Factory Mapping Data Flow

KROMERM. **Mapping data flows - Azure Data Factory.** Microsoft.com. Disponível em:

<<https://docs.microsoft.com/en-us/azure/data-factory/concepts-data-flow-overview>>. Acesso em: 12 May 2021.

Azure Analysis Services

MINEWISKAN. **What is Azure Analysis Services.** Microsoft.com. Disponível em:

<<https://docs.microsoft.com/en-us/azure/analysis-services/analysis-services-overview>>.

Acesso em: 12 May 2021.

Azure SQL Database

STEVESTEIN. **What is the Azure SQL Database service? - Azure SQL Database.**

Microsoft.com. Disponível em: <<https://docs.microsoft.com/en-us/azure/azure-sql/database/sql-database-paas-overview>>. Acesso em: 12 May 2021.

DOS SANTOS, A. F. P. **AVA Login | UniCarioca.** Disponível em:

<https://ava.unicarioca.edu.br/graduacao/pluginfile.php/969791/mod_resource/content/0/Tema%206%20-%20Resumo.pdf>. Acesso em: 5 maio. 2021.

ZOINERTEJADA. **Processando arquivos CSV e JSON - Azure Architecture Center.**

Disponível em: <<https://docs.microsoft.com/pt-br/azure/architecture/data-guide/scenarios/csv-and-json>>. Acesso em: 22 maio. 2021.

Visualizadores de dados

Guia prático da visualização de dados: definição, exemplos e recursos de aprendizado.

Tableau. Disponível em: <<https://www.tableau.com/pt-br/learn/articles/data-visualization>>.

Acesso em: 3 May 2021.

MIHART. **O que é Power BI? - Power BI.** Microsoft.com. Disponível em:

<<https://docs.microsoft.com/pt-br/power-bi/fundamentals/power-bi-overview>>. Acesso em: 3 May 2021.

Power Query

WEBB, C. **Power Query for Power BI and Excel.** [s.l.] Berkeley, Ca Apress, 2014.

DAX

MINEWISKAN. **Referência do DAX (Data Analysis Expressions) - DAX.** Disponível em:

<<https://docs.microsoft.com/pt-br/dax/>>. Acesso em: 13 maio. 2021.

What is DAX - DATA ANALYSIS EXPRESSIONS - PowerPivot & PowerBI. Disponível em: <<https://theexcelclub.com/what-is-dax-data-analysis-expressions/>>. Acesso em: 13 maio. 2021.

Azure Storage Explorer – cloud storage management | Microsoft Azure. Microsoft.com. Disponível em: <<https://azure.microsoft.com/en-us/features/storage-explorer/>>. Acesso em: 10 May 2021.

What Is an IDE? | Codecademy. Codecademy. Disponível em: <<https://www.codecademy.com/articles/what-is-an-ide>>. Acesso em: 10 May 2021.

RStudio, new open-source IDE for R. Rstudio.com. Disponível em: <<https://blog.rstudio.com/2011/02/28/rstudio-new-open-source-ide-for-r/>>. Acesso em: 30 Apr 2021.

TERRYGLEE; OLPROOD. Visão geral do Visual Studio. Disponível em: <<https://docs.microsoft.com/pt-br/visualstudio/get-started/visual-studio-ide?view=vs-2019>>. Acesso em: 5 maio. 2021.

MINEWISKAN; OLPROOD. Ferramentas de Analysis Services. Disponível em: <<https://docs.microsoft.com/pt-br/analysis-services/tools-and-applications-used-in-analysis-services?view=asallproducts-allversions>>. Acesso em: 5 maio. 2021.

YUALAN. Download and install Azure Data Studio - Azure Data Studio. Microsoft.com. Disponível em: <<https://docs.microsoft.com/en-us/sql/azure-data-studio/download-azure-data-studio?view=sql-server-ver15>>. Acesso em: 10 May 2021.

SQL SERVER MANAGEMENT STUDIO

DZSQUARED. Download SQL Server Management Studio (SSMS) - SQL Server Management Studio (SSMS). Microsoft.com. Disponível em: <<https://docs.microsoft.com/en-us/sql/ssms/download-sql-server-management-studio-ssms?view=sql-server-ver15>>. Acesso em: 13 May 2021.

OLIVEIRA, Marcos. O que é SGBD ? Terminal Root - Linux e Desenvolvimento. Disponível em: <<https://terminalroot.com.br/2019/08/o-que-e-sgbd.html>>. Acesso em: 10 May 2021.

Data de acesso 30/04/2021

<https://www.r-project.org/about.html>

The Comprehensive R Archive Network. R-project.org. Disponível em: <<https://cran.r-project.org/>>. Acesso em: 11 May 2021.

PRESCOTT, P. **SQL Para Iniciantes**. Tradução: Rafaela C. S. Barros. [s.l.] Babelcub, Inc, 2015.

ROVEDA, U. **SQL: o que é, para que serve e como aprender comando SQL**. Disponível em: <<https://kenzie.com.br/blog/sql/>>. Acesso em: 5 maio. 2021.

UNIVERSITY OF WARWICK, COVENTRY, UK. The R User Conference, useR! 2011. [s.l: s.n.].

VERZANI, J. Getting Started with RStudio. Sebastopol: O'Reilly Media, Inc, 2011. p. 3–4
Desenvolvimento

NOÉ, M. **Média Aritmética**. Disponível em:

<<https://brasilecola.uol.com.br/matematica/media-aritmetica.htm>>. Acesso em: 27 maio. 2021.

Criar um gráfico de mapa do Office. Disponível em: <<https://support.microsoft.com/pt-br/office/criar-um-gr%C3%A1fico-de-mapa-do-office-dfe86d28-a610-4ef5-9b30-362d5c624b68>>. Acesso em: 27 maio. 2021.

Criar um gráfico de cascata. Disponível em: <<https://support.microsoft.com/pt-br/office/criar-um-gr%C3%A1fico-de-cascata-8de1ece4-ff21-4d37-acd7-546f5527f185#:~:text=Um%20gr%C3%A1fico%20de%20cascata%20mostra,de%20valores%20positivos%20e%20negativos.>>. Acesso em: 27 maio. 2021.

Criar Um Gráfico De Funil. Disponível em: <<https://support.microsoft.com/pt-br/office/criar-um-gr%C3%A1fico-de-funil-ba21bcba-f325-4d9f-93df-97074589a70e>>. Acesso em: 27 maio. 2021.

MIHART; OLPROD. **Árvore De Decomposição - Power BI**. Disponível em: <<https://docs.microsoft.com/pt-br/power-bi/visuals/power-bi-visualization-decomposition-tree>>. Acesso em: 27 maio. 2021.

SANTOS, D. **Pesquisa sobre o entendimento da desigualdade social através da mineração aplicada a base de dados do ENEM 2019**, 27 maio 2021.