Aplicação de “Business Intelligence” e “Data Visualization” nos indicadores de desempenho da educação básica brasileira usando a metodologia CRISP-DM

Charles Fernandes Cardoso Júnior[1]*; Ricardo Limongi[2]

[1]Universidade Salgado de Oliveira. Analista de Sistemas. Av Cora Coralina, Q F25 L40, S. Sul - CEP 74.080-445 Goiânia, Goiás, Brasil
[2] Doutor em Administração. Rua 74, n 240. Goiânia-GO. - CEP 74810-380.
Goiânia, Goiás, Brasil.
*autor correspondente: charlesjuniorx@gmail.com

Aplicação de “Business Intelligence” e “Data Visualization” nos indicadores de desempenho da educação básica brasileira usando a metodologia CRISP-DM

Resumo

Um dos fatores que contribuem para o desenvolvimento de um país e o bem estar de seu povo é a educação. No Brasil, o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira [INEP] avalia o desempenho escolar e disponibiliza os principais indicadores da educação básica. O presente trabalho pesquisou técnicas e criou um ambiente centralizado que possibilitou a análise para as tomadas de decisões por meio da metodologia “Cross-Industry Standard Process for Data Mining [CRISP-DM]”, “Business Intelligence [BI]” e “Data Visualization [DATAVIZ]”. A principal contribuição da pesquisa é permitir a facilidade em futuros projetos com a base INEP envolvendo BI e DATAVIZ.
Palavras-chave: análise; modelo; decisão; ensino; estruturação.

Introdução

No Brasil existe uma política de levantamento e disponibilização dos dados educacionais que visam contribuir com as ações empregadas na melhoria do ensino nos níveis existentes (Fonseca e Namen, 2016). A instituição responsável pela obtenção e publicação desses dados é o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira [INEP] e contém informações sobre as etapas do ensino da educação básica (infantil, fundamental e médio) e de nível superior (Nascimento et al., 2018).

O INEP realiza pesquisas e avaliações dos conteúdos ensinados pelo sistema de educação, além de diversas questões que possam interferir na qualidade do ensino e desempenho dos estudantes (Fonseca e Namen, 2016) como adequação da formação do docente, complexidade de gestão da escola, média de horas-aula diária, docentes com curso superior, etc (INEP, 2022). O INEP disponibiliza em sua plataforma os resultados de avaliações ocorridas em um ano específico (o que torna custosa a verificação da evolução ao longo dos anos), além de estarem de forma misturada entre os tipos da educação básica existentes e sem o uso de recursos visuais como gráficos de barras ou linhas, mapas ou valores em “cards” (INEP, 2022). Os dados disponibilizados pelo INEP já foram usados para o desenvolvimento de vários trabalhos: mineração de dados educacionais, desempenho dos alunos, fatores que influenciam a aprendizagem, etc (Nascimento et al., 2018). Seria interessante criar uma estrutura para a disponibilização destes dados que permita a realização de análises e acompanhamentos pelos órgãos envolvidos e partes interessadas (por exemplo, gestores educacionais nas esferas: federal, estadual, municipal, unidade escolar).

As fases do “Cross-Industry Standard for Data Mining [CRISP-DM]” são empregadas em processos de mineração de dados e são bastante utilizadas devido a sua flexibilidade e facilidade de implementação (Huber et al., 2019). Foram aplicadas no trabalho as seis etapas que visam o entendimento do negócio, compreensão dos dados e como estes devem ser

preparados para serem utilizados durante a criação de modelos elaborados com critérios pré-definidos (Colpani, 2018). Após a criação dos modelos, foram realizadas avaliações para identificar se seriam ou não disponibilizados para o acesso aos recursos criados (Plotnikova et al., 2022).

"Business Intelligence [BI]" envolve teorias, metodologias, processos, tecnologias, etc, para compor uma estrutura onde os dados são trabalhados para prover vários "insights" que irão ajudar na gestão e no suporte às tomadas de decisões (Nery, 2013). Existem várias técnicas envolvidas na implementação de um projeto de BI e dentre elas foi implementada no trabalho a criação de uma estrutura para o armazenamento dos dados conhecida como "Data Warehouse [DW]", com um modelo de dados multidimensional composto por tabelas de dimensões, tabelas fatos e várias outras técnicas correlacionadas com a perspectiva da visualização (Kimball e Ross, 2013).

"Data Visualization [DATAVIZ]" é a apresentação dos dados por um conjunto de recursos visuais e bem estruturados, como gráficos, mapas, "cards", esquemas, "storytelling", etc. Foi usado através da aplicação de técnicas de visualização bem fundamentadas que proporcionaram uma comunicação mais clara e objetiva entre os dados apresentados e as análises realizadas (Sadiku et al., 2016).

O presente estudo buscou aplicar modelos, técnicas e conceitos que melhorem a visualização dos indicadores da educação básica brasileira através da criação de uma proposta para a otimização no uso dos dados fornecidos pelo INEP (Chawla et al., 2018). Para isso, utilizou-se a metodologia CRISP-DM, com suas etapas flexíveis, visando o auxílio na elaboração, desenvolvimento e condução do projeto (Huber et al., 2019). Foram empregadas as técnicas existentes para a limpeza, extração, preparação e carga de dados (usando a linguagem R e o "Power" BI). Foi criado um modelo multidimensional aplicando as técnicas utilizadas em projetos de BI (Nery, 2013). Foram pesquisadas e aplicadas técnicas de visualização de dados onde a disponibilização destes evoluiu para além de sua forma pobre em recursos visuais, o que possibilitou melhorias nas análises e o entendimento sobre os dados (Knaflic, 2019).

A partir da estruturação e conceitos abordados neste trabalho, que serve como apoio à tomada de decisão, foi disponibilizada uma proposta para a aplicação de uma metodologia simples e de fácil implementação que contribua para uma melhoria na apresentação e análise dos dados envolvidos (Chawla et al., 2018). Uma vez que a estrutura apresentada seja desenvolvida e disponibilizada, os gestores educacionais federais, estaduais, municipais, etc, podem fazer uso dessa estrutura para analisarem as avaliações de desempenho escolar, fornecidos pelo INEP, de forma rápida, fácil, segmentada e com múltiplas visões acerca dos indicadores educacionais.

Materiais e métodos

Na condução do desenvolvimento do trabalho, foi usada a metodologia CRISP-DM, que é considerada ser bastante popular e com uso bem difundido entre as empresas (Huber et al., 2019). Trata-se de uma importante ferramenta que contribui com o sucesso dos processos relacionados à mineração de dados e em sintonia com os objetivos organizacionais, sendo ajustável conforme a necessidade (Plotnikova et al., 2022). Por conter uma sequência de etapas flexíveis para a construção de um modelo real de mineração de dados, essa metodologia ajudou nas questões de negócio, preparação e apresentação dos modelos (Nascimento et al., 2018). Na Figura 1 são apresentadas as fases existentes na metodologia CRISP-DM e subsequentemente, cada uma dessas fases são discutidas conforme foram sendo aplicadas durante todo o desenvolvimento do trabalho. Como a proposta para a criação da estrutura de visualização dos dados do INEP envolveram três pilares, os quais são CRISP-DM, BI e DATAVIZ, é importante o entendimento que BI e DATAVIZ foram sendo apresentados e aplicados dentro de cada uma dessas fases do CRISP-DM, conforme o seu desenvolvimento, e que as características desses três pilares foram discutidas em paralelo ao longo da composição do trabalho e de acordo com a sequência existente em cada etapa.

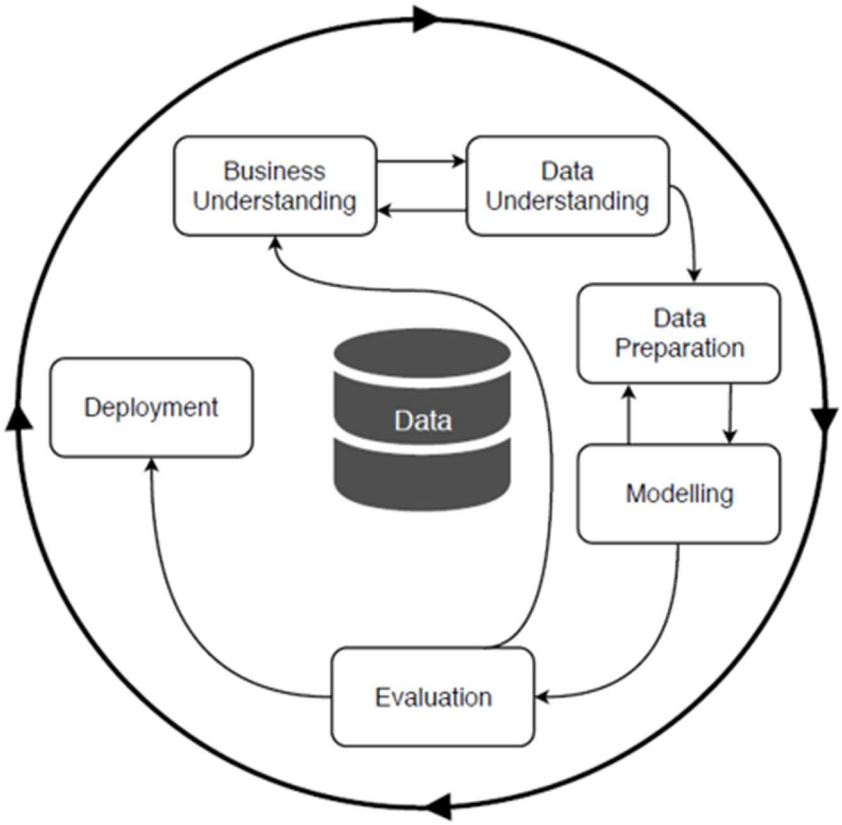

Figura 1. Fases do CRISP-DM
Fonte: Martínez et al. (2019)

“Business Understanding”

Segundo Colpani (2018), o CRISP-DM se inicia na fase de compreensão do negócio com uma contextualização do que será realizado. A intenção é levantar as informações necessárias para o entendimento do negócio, buscando otimizar e maximizar a eficiência do desenvolvimento de soluções através da mineração de dados (Huber et al., 2019). Foi realizada uma busca por informações sobre o cenário educacional em trabalhos correlacionados e também por documentação no órgão responsável pelos dados e estatísticas educacionais no Brasil (Nascimento et al., 2018). Foram compreendidas algumas questões inerentes ao INEP, tais como: elaboração dos dados, metodologia das avaliações, formas de disponibilização dos dados, possíveis cenários de aplicabilidade dos dados com a finalidade de tomada de decisões, etc (Colpani, 2018). Após compreender o contexto de negócio envolvido, iniciaram-se alguns estudos preliminares sobre como este contexto poderia ser aplicado em cima de uma estrutura tecnológica voltada para uma aplicação de BI e DATAVIZ.

O CRISP-DM permite uma abordagem bastante flexível (Laureano et al., 2014), por exemplo, nas fases de compreensão do negócio e compreensão dos dados, houve um certo ir e vir entre essas etapas para refinar o entendimento e fazer os ajustes necessários (Huber et al., 2019). Nessa fase fez-se o uso das técnicas de BI para identificar quais os tipos de análises e personagens tinham interesse sobre os dados do INEP, os quais foram classificados de acordo com a localidade e o contexto de consultas a serem realizadas nesses locais. A Tabela 1 resume quais foram esses personagens identificados e quais as análises pretendidas por estes.

Tabela 1: Personagens e análises pretendidas

Personagem	Análises pretendidas
Gestor educacional federal	Análise sobre os indicadores relacionados a todas as escolas brasileiras.
Gestor educacional estadual	Análise sobre os indicadores relacionados às escolas de seu estado.
Gestor educacional municipal	Análise sobre os indicadores relacionados às escolas de seu município.
Gestor educacional da escola	Análise sobre os indicadores relacionados à uma escola em específico.

Fonte: Dados originais da pesquisa

Segundo Provost e Fawcett (2016) a tomada de decisão orientada por dados é justamente estabelecer uma prática de decisões na análise dos dados e não apenas na experiência ou intuição. A Tabela 2 apresenta quais as visões e as análises que os gestores

educacionais desejam obter sobre os indicadores educacionais e que devem ser garantidas a existência no modelo multidimensional e visual (“dashboards”).

Tabela 2: Tipos de visão sobre os dados

Visão	Descrição
Sobre o desempenho escolar	Apresentar as taxas do desempenho escolar para aprovação, reprovação e abandono, juntamente com os indicadores de hora-aula diária, alunos por turma, regularidade do docente e distorção idade-série e categoria.
Sobre a evolução do desempenho escolar	Apresentar a evolução temporal da taxa de desempenho escolar dos últimos 5 anos, onde os dados para aprovação e reprovação estejam juntos para comparação e a taxa de abandono esteja separado destes.
Sobre a categoria e dependência administrativa	Apresentar a evolução temporal dos últimos 5 anos, apenas para a taxa de aprovação relacionada aos dados de dependência administrativa e de categoria.
Sobre a complexidade de gestão escola	Apresentar a quantidade de escolas contidas em cada nível de complexidade juntamente com os valores das taxas de aprovação. Apresentar também os valores de aprovação, reprovação e abandono para todos os níveis de complexidade.

Fonte: Dados originais da pesquisa

“Data Understanding”

Nessa fase foi acessada a plataforma do INEP a fim de obter os dados que estavam disponíveis para “download” (INEP, 2022). Foram realizadas verificações para entender o formato dos dados e testes que permitiram familiarizar-se e evidenciar a existência de possíveis problemas de qualidade (Nascimento et al., 2018). Foi identificado que os dados são fornecidos através de vários arquivos de acordo com o assunto e ano, e que seria necessário baixar cada um desses arquivos para a aplicação ao desenvolvimento do trabalho.

A Tabela 3 mostra quais foram os arquivos baixados do INEP. Na tabela existem três colunas. A primeira coluna contém o nome original do arquivo de acordo com a plataforma do INEP (sem o sufixo ano). Neste trabalho foram usados os arquivos com o sufixo “_2017” até “_2021”, ou seja, cada grupo contém cinco arquivos com os dados dos anos de 2017 até 2021 (no início deste trabalho ainda não tinha sido disponibilizado o arquivo com os dados de desempenho escolar (TXR) para o ano de 2022, então optou-se por trabalhar até o ano de 2021). No total, foram usados 30 arquivos do INEP. A segunda coluna contém uma descrição sobre qual é o tipo de informação contida no arquivo. Na terceira coluna está uma definição

da sigla adotada por este trabalho (baseado no prefixo do nome do arquivo) para identificar o arquivo e seus dados. Com essa definição, o trabalho passou a fazer uso de siglas para referenciar os arquivos baixados, onde cada grupo de cinco arquivos foram consolidados em um único arquivo, que representa o grupo, e então cada arquivo de grupo foi consolidado em um único arquivo geral (contendo todos os dados necessários). Ainda nessa fase, foram levantadas quais as colunas que deveriam ser removidas dos arquivos pela atividade de "data wrangling" (limpeza e preparação dos dados) e de extração, transformação e carga [ETL]. Observou-se que as avaliações aplicadas pelo INEP foram sobre todos os níveis da educação básica, os quais são: infantil, fundamental e médio. Todavia, nessa fase, foi estabelecido que este trabalho atuaria apenas com os dados do ensino fundamental e médio, pois são os únicos existentes em todos os arquivos envolvidos. Porém, isso não impede que o trabalho seja adaptado, futuramente, para conter todos os níveis existentes na educação básica.

Tabela 3: Arquivos baixados do INEP

Nome do arquivo (sem sufixo)	Descrição do arquivo	Sigla
tx_rend_escolas.xlsx	Taxa de Rendimento Escolar	TXR
atu_escolas.xlsx	Média de Alunos por Turma	ATU
had_escolas.xlsx	Média de Horas-Aula Diária	HAD
tdi_escolas.xlsx	Taxa de Distorção Idade-Série	TDI
icg_escolas.xlsx	Indicador de Complexidade Gestão Escola	ICG
ird_escolas.xlsx	Indicador de Regularidade do Docente	IRD

Fonte: Dados originais da pesquisa

"Data Preparation"

É nessa fase que o CRISP-DM emprega os esforços de limpeza, ajustes e preparação dos dados (Laureano et al., 2014). Os arquivos do INEP continham algumas formatações inadequadas de estilos, cabeçalhos e rodapés. Como os dados da educação básica estavam misturados, foi preciso fazer um ajuste para deixar apenas aqueles relacionados aos ensinos médio e fundamental (Nascimento et al., 2018). Existem atualmente diversas abordagens e ferramentas tecnológicas que auxiliam nas atividades de "data wrangling" e de ETL.

Para Diamond e Mattia (2017) a visualização dos dados é um componente chave na análise e que existem vários recursos tecnológicos para essa finalidade. Este trabalho optou pelo uso da linguagem R e da estrutura de preparação de dados do "Power" BI. Essa escolha foi motivada pela característica do trabalho e natureza dos dados envolvidos (os arquivos depois de disponibilizados pelo INEP praticamente não sofrem alterações significativas). Tal cenário tem um comportamento diferente das aplicações de BI criadas em cima de um sistema de gestão usado por uma indústria ou comércio, onde o negócio e os dados de origem sofrem alterações diárias e por isso o processo de ETL requer uma maior força nas atividades que

envolvem a obtenção, tratamento e carga de dados, sendo necessária a criação de uma estrutura de DW mais robusta e um uso mais amplo das técnicas existentes (Kimball e Ross, 2013).

A Figura 2 resume como foi feito o processo de preparação dos dados. Após serem baixados os 30 arquivos da plataforma do INEP (em formato do Excel), foi realizado um processo de limpeza e preparação dos arquivos, usando a linguagem R e os pacotes "tydeverse" e "readxl". Nesse momento foram removidos os cabeçalhos e rodapés existentes nos arquivos, foram retiradas todas as colunas que não seriam utilizadas, deletados vários registros (será explicado o motivo dessa deleção mais adiante), e por fim foi criado um arquivo único contendo todos os dados necessários para a aplicação do DW junto ao trabalho (no "Power" BI, esse arquivo único foi utilizado para a criação do modelo multidimensional). Assim como ocorreu nas fases de compreensão do negócio e dos dados, as fases de preparação de dados e de modelagem também operaram de forma paralela, pois durante a criação do modelo tiveram dados que precisaram ser adicionados e outros que foram ajustados (Laureano et al., 2014).

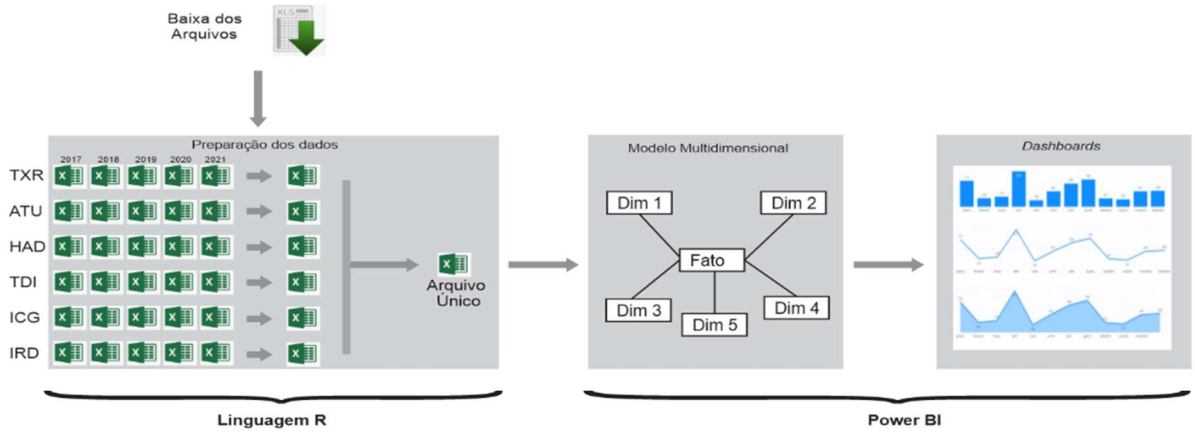

Figura 2. Preparação dos dados, criação do modelo dimensional e visualização
Fonte: Dados originais da pesquisa

"Modelling"

Nesta fase do CRISP-DM foi criado o modelo multidimensional conforme as técnicas de BI e o modelo visual conforme o DATAVIZ. É apresentado abaixo os detalhes que envolveram a criação do modelo multidimensional e logo após serão apresentados os detalhes envolvidos na criação do modelo visual. Segundo Nery (2013), há duas abordagens para a criação de um modelo multidimensional: "star schema" e "snowflake". No "Star schema" dimensões se relacionam apenas com fato e por isso são mais performáticas, já no "snowflake" dimensões se relacionam também com dimensões. Neste trabalho foi aplicado o modelo "star schema".

Existem dois tipos de tabelas: Dimensões e Fatos. As tabelas de dimensões, contém a descrição ou rótulo do fato acontecido (Kimball e Ross, 2013). As tabelas fatos contém as medidas com valores quantitativos (discretos ou contínuos) e como o próprio nome diz, representa um fato ocorrido em um negócio (Kimball e Ross, 2013). Por exemplo: A prova feita por um aluno é um fato, pois é uma ação ou acontecimento que ocorreu e este não pode mais ser mudado. Fatos são usados em BI para entender o que ocorreu e assim tomar as melhores decisões. Essa combinação de dimensão e fato permitiu a criação da engrenagem para a visualização dos dados do INEP sobre várias perspectivas, como em um cubo mágico. Foram usadas no trabalho três tipos de dimensões dentre as várias existentes. Segundo Nery (2013), na construção de um DW, as dimensões buscam responder quatro perguntas em relação à ocorrência de um fato: onde ocorreu o fato, quando ocorreu o fato, quem está envolvido no fato e o que está envolvido no fato. Uma mesma dimensão pode responder a mais de uma das perguntas apresentadas, dependendo do tipo de visão dos dados que está em análise (Nery, 2013).

Neste trabalho, devido à necessidade de mostrar os dados relacionados ao período de 2017 até 2021, foi necessária a criação de uma dimensão de tempo (Kimball e Ross, 2013). Existe nos arquivos do INEP uma coluna que indica o ano correspondente aos dados disponibilizados. Durante a fase de preparação dos dados foi criada uma nova coluna do tipo “date” com a junção dos valores fixos “31/12/” e o respectivo ano (Lachev, 2022). Por exemplo, para o ano de 2017 os valores nessa nova coluna ficaram como: 31/12/2017. A ferramenta de visualização dos dados (“Power” BI) trabalha com o tipo “date” e por isso é necessário fornecer a data em um formato válido pela ferramenta (Lachev, 2022). A existência de uma dimensão de tempo, permite que seja feita uma avaliação temporal do desempenho das escolas durante o período de cinco anos (solicitado na fase de negócio) e responde à pergunta sobre quando ocorreram os fatos analisados pelo INEP (Nery, 2013).

Foi criada uma dimensão hierárquica com as informações de localidade dos dados (região, estado, município e escola) e respondem à pergunta sobre onde ocorreu o fato (Nery, 2013). As demais dimensões aplicadas neste trabalho não têm características em especial e se encaixam em uma dimensão “slowly changing” do tipo um. Todas as dimensões podem ser consideradas como sendo uma “slowly changing” do tipo um ou do tipo dois. A diferença é que quando ocorre alteração na origem, os dados são apenas atualizados na “slowly changing” do tipo um e são inseridos novamente na do tipo dois fazendo um controle sobre qual é o registro atual (Kimball e Ross, 2013). Em um modelo multidimensional, existem várias técnicas onde o entendimento é indispensável para a criação das tabelas de dimensões e de fatos, tais como: granularidade, “surrogate key” e as chaves primárias e estrangeiras (Kimball e Ross, 2013).

Chaves primárias e estrangeiras são bastante usadas em modelos de banco de dados relacionais e aplicadas em um modelo multidimensional para fazer uma ligação entre os registros deste com os registros do modelo externo (origem dos dados) e também para estabelecer o relacionamento entre dimensão e fato (Nery, 2013). Nesse contexto, durante o processo de ETL é possível identificar se o registro que está sendo carregado é uma inserção ou alteração e fazer a ação necessária. Chaves primárias dos dados externos (origem) tornam-se colunas comuns em um modelo multidimensional, pois não é aplicada regras de integridade referencial entre os modelos e quem garante esse relacionamento é o processo de carga dos dados (Nery, 2013).

A “surrogate key” tem o mesmo conceito de garantia da integridade referencial das chaves primárias e estrangeiras em um modelo relacional, porém são aplicadas para relacionar dimensões e fatos (Kimball e Ross, 2013). Por questão de convenção de nomenclatura, este trabalho adotou o prefixo sk_ para a “surrogate key” e o prefixo id_ para a coluna que é o identificador de registro entre a origem (modelo externo) e o modelo multidimensional. Na tabela fato todas as “surrogates keys” fazem parte de uma chave primária composta para evitar a duplicidade de um mesmo fato (Kimball e Ross, 2013).

A granularidade é uma técnica com bastante impacto na criação de um modelo multidimensional porque pode limitar o nível de detalhamento dos dados (interferindo na análise e tomada de decisão) e causar problemas de lentidão durante as consultas, devido ao enorme volume de dados que poderá existir em uma granularidade com maior detalhamento (Nery, 2013). Quanto menor o grão do modelo multidimensional, maior o nível de detalhamento e menor é a performance. Quanto maior o grão, menor será o nível de detalhamento e maior será a performance. Esse trabalho usou a dimensão hierárquica com os dados de localidade para a aplicação do nível de granularidade (Nery, 2013). A Figura 3 resume como seria a visualização sobre os dados de acordo com uma alteração na granularidade do trabalho. Um nível de detalhamento baixo poderia impedir a visão sobre os dados para alguns gestores educacionais, como por exemplo o municipal e o de unidade escolar.

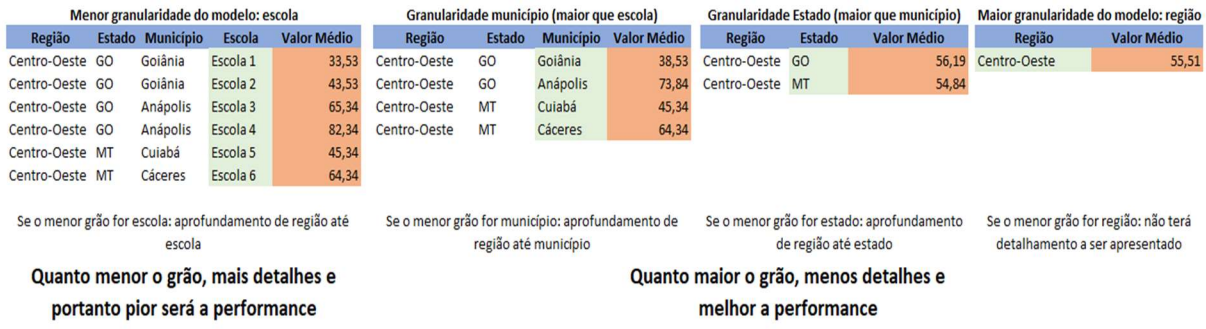

Figura 3. Granularidades possíveis de serem aplicadas neste trabalho
Fonte: Dados originais da pesquisa

É a granularidade que permite aplicar as técnicas de “Drill Down” e de “Roll Up” em componentes de visualização de dados, como os gráficos do “Power” BI. “Drill Down” significa aumentar o nível de detalhe, ou seja, ir para o menor grão. “Roll Up” significa diminuir o nível de detalhe, ou seja, subir para o maior grão (Kimball e Ross, 2013). Como dito, o menor grão aplicado nos dados do INEP foi a escola e caso tenha um gráfico de barras com as métricas de região, após fazer o “Drill Down” será apresentado apenas os dados dos estados pertencentes à região escolhida e depois apenas as cidades pertencentes ao estado escolhido e por fim, apenas as escolas da cidade escolhida, que é o último grão e contém o maior nível de detalhamento. A agregação dos dados de acordo com a granularidade aplicada ao modelo é feita durante o processo de ETL (Nery, 2013). Neste trabalho não foi necessário fazer tal agregação. Se fossem usados os microdados da educação básica (INEP, 2022) que contém os detalhes dos alunos, a escola deixaria de ser o menor grão e este passaria a ser o aluno (se fosse a necessidade do negócio).

No modelo multidimensional deste trabalho, foi criada apenas uma tabela fato. Ficou decidido pelo negócio que a carga de dados traria para o modelo apenas os dados que existissem simultaneamente na granularidade mais baixa em todos os arquivos envolvidos (Nery, 2013). Obviamente em um cenário onde o negócio necessite uma abordagem diferente, deverá ser realizada as devidas mudanças para atender aos requisitos que forem necessários (Kimball e Ross, 2013). Por exemplo, caso fosse estabelecido pelo negócio uma análise onde tivesse que ter todos os dados, independente da referida simultaneidade, o modelo teria que ser adaptado e poderia ter até seis tabelas fatos, uma para cada arquivo carregado. O que mudaria todo o processo de preparação, carga e criação do modelo. É bastante perceptível que um modelo multidimensional é criado para atender à necessidade estabelecida pelo negócio (Nery, 2013). Uma tabela fato contém basicamente as “surrogates keys” das dimensões ligadas a esta (que devem garantir a existência única de um fato, ou seja, não deverá existir mais de um fato referente ao mesmo acontecimento) e os valores a serem medidos (métricas) pelo negócio (Kimball e Ross, 2013). Esse trabalho fez o uso de uma tabela fato do tipo transacional, onde cada transação na origem representa um fato.

A Figura 4 mostra como o modelo multidimensional foi criado para atender às necessidades elencadas pelo negócio e para o auxílio nas tomadas de decisões e aplicando todas as técnicas de BI aqui apresentadas (visando a aplicação das técnicas de DATAVIZ*)*. Foram criadas seis tabelas de dimensões e apenas uma tabela fato. A dimensão D_Dependencia originou-se da coluna dependência administrativa que contém os valores: federal, estadual, municipal e privada. A dimensão D_Categoria originou-se da coluna que categoriza uma escola em rural ou urbana. A dimensão D_Local foi criada a partir da hierarquia de localização, existente entre as colunas região, estado, município e escola (sendo escola o menor grão do modelo). A dimensão D_Periodo foi criada a partir da coluna ano. A

dimensão D_Nivel originou-se da coluna com os níveis de complexidade da gestão de escola (do nível um ao nível seis, conforme INEP). A dimensão D_Ensino é a classificação das taxas de rendimento escolar em ensino fundamental e médio. A tabela fato F_Desempenho é composta pelas colunas com os valores em todos os arquivos envolvidos e as "surrogates keys" que a relaciona com as dimensões (por convenção, identificadas com um prefixo sk_).

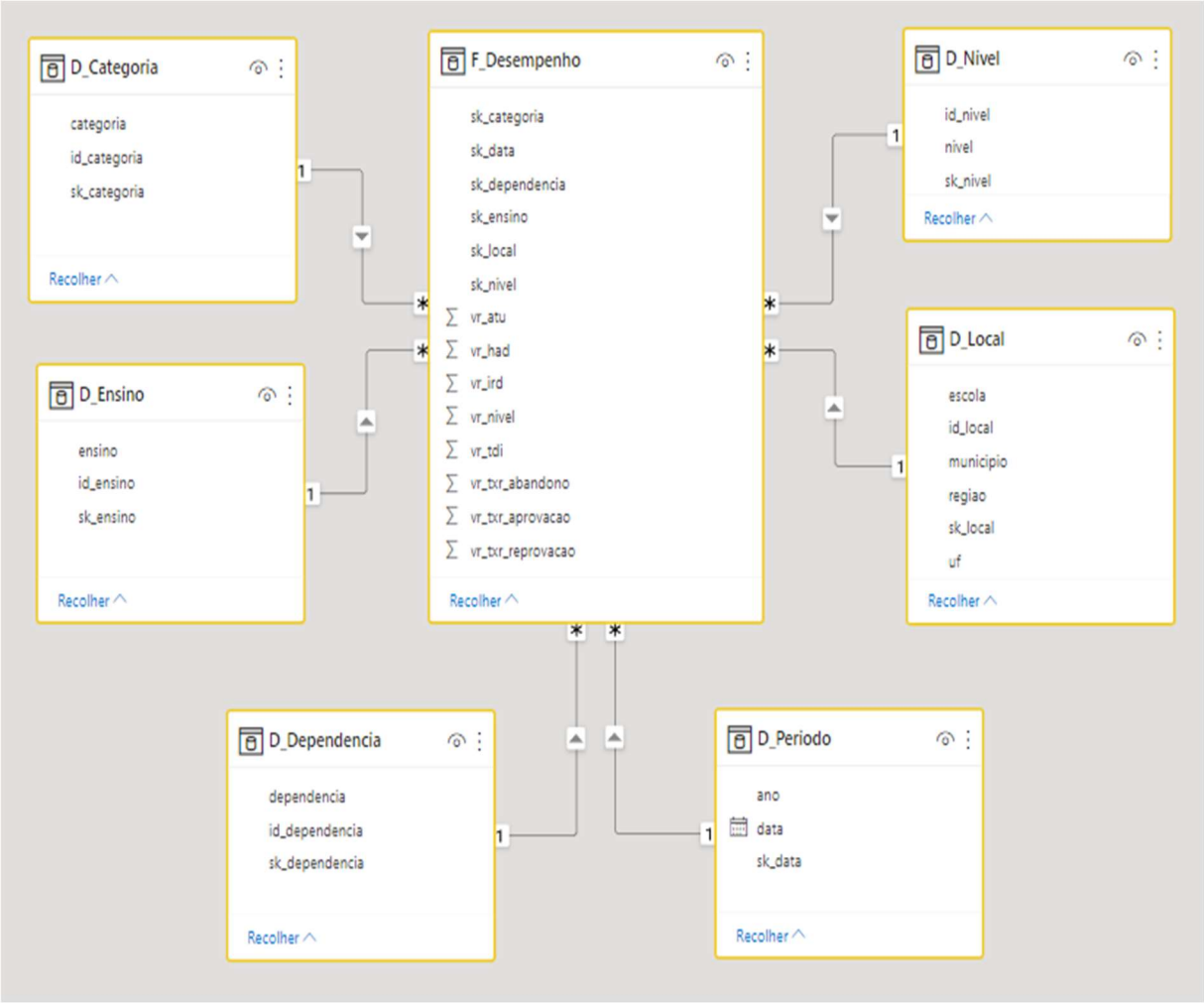

Figura 4. Modelo multidimensional aplicado no trabalho
Fonte: Dados originais da pesquisa

Foram criados quatro modelos visuais para atenderem às necessidades dos gestores educacionais que foram levantadas na fase de entendimento de negócio. As cores aplicadas em todos os gráficos seguiram as indicações para daltonismo e foi aplicado recursos de "storytelling" com dados (Knaflic, 2019). Segundo Silva (2019), uma abordagem visual limpa, significa um conjunto de dados no formato apropriado para a maioria das ferramentas de visualização e as representações visuais construídas nesse trabalho seguiram por esse caminho. A Figura 5 mostra as quatro telas criadas nessa fase do CRISP-DM com as técnicas de DATAVIZ e que serão apresentadas individualmente na próxima etapa que aborda sobre a fase de avaliação dos modelos criados.

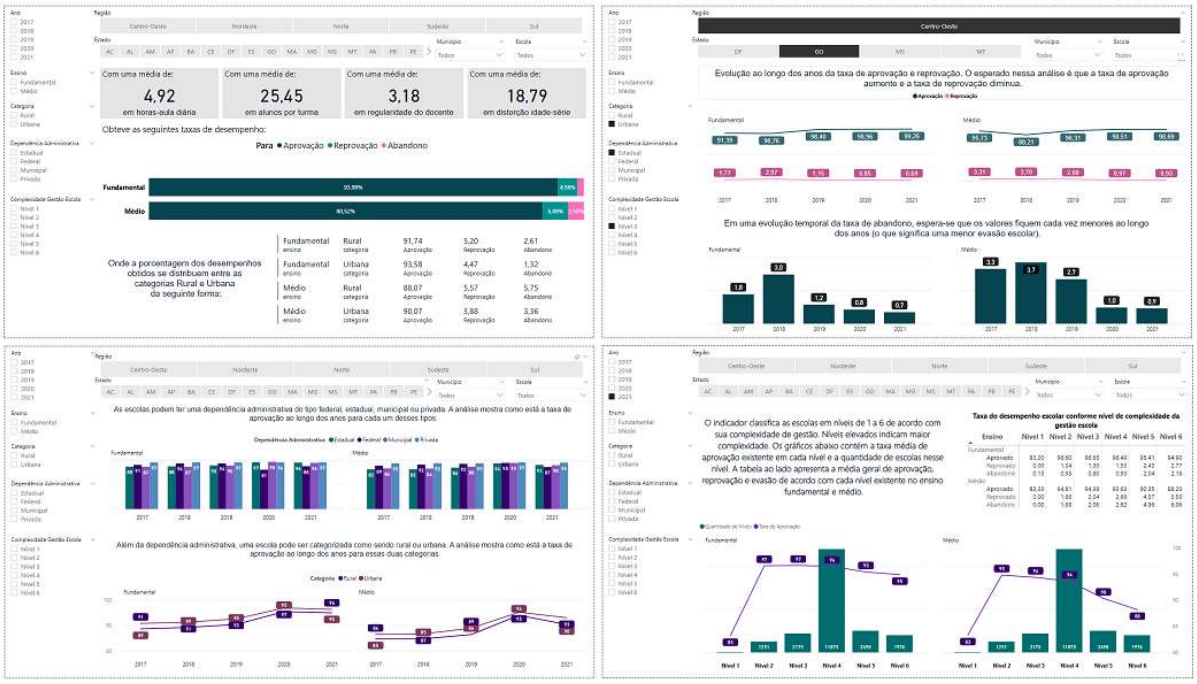

Figura 5. Modelos visuais criados aplicando DATAVIZ
Fonte: Dados originais da pesquisa

Resultados e Discussão

“Evaluation”

Essa é a etapa do CRISP-DM em que os modelos criados são avaliados para verificar se o resultado corresponde à expectativa do projeto (Huber et al., 2019). Se não estiver de acordo com o desejado ou caso exista uma consideração que leve a um entendimento que deva aplicar uma melhoria, então deverá empregar as ações necessárias aos ajustes (Plotnikova et al., 2022). Nessa fase do CRISP-DM, foram avaliados os resultados apresentados pelos modelos criados, através da aplicação de várias segmentações visando atender ao que foi levantado na fase de entendimento do negócio.

A Figura 6 mostra os dados do desempenho escolar com um resultado voltado para um gestor nacional sem segmentação. Foi trabalhado o conceito de "storytelling" com dados em que é possível, por exemplo, obter a seguinte história sobre os dados (relacionados à taxa de aprovação): “Com uma média de 4,92 em horas-aula diária, obteve-se uma taxa de aprovação de 93,99% para o ensino fundamental e de 90,52% para o ensino médio, onde a porcentagem dos desempenhos obtidos se distribuem entre as categorias rural e urbana da seguinte forma: fundamental rural 91,74%, fundamental urbana 93,58%, médio rural 88,07% e médio urbana 90,07%, relacionados à aprovação” (Knaflic, 2019). Obviamente a história contada acima trata-se da análise de apenas uma dentre as várias possibilidades existentes na tela, como: “Com uma média de 25,45 alunos por turma e 18,79 em distorção idade-série,

obteve-se uma taxa de reprovação de 4,56% para o ensino fundamental e de 5,89% para o ensino médio, onde a porcentagem dos desempenhos obtidos se distribuem entre as categorias rural e urbana da seguinte forma: fundamental rural 5,20%, fundamental urbana 4,47%, médio rural 5,57% e médio urbana 5,88%, relacionados à reprovação" (Knaflic, 2019).

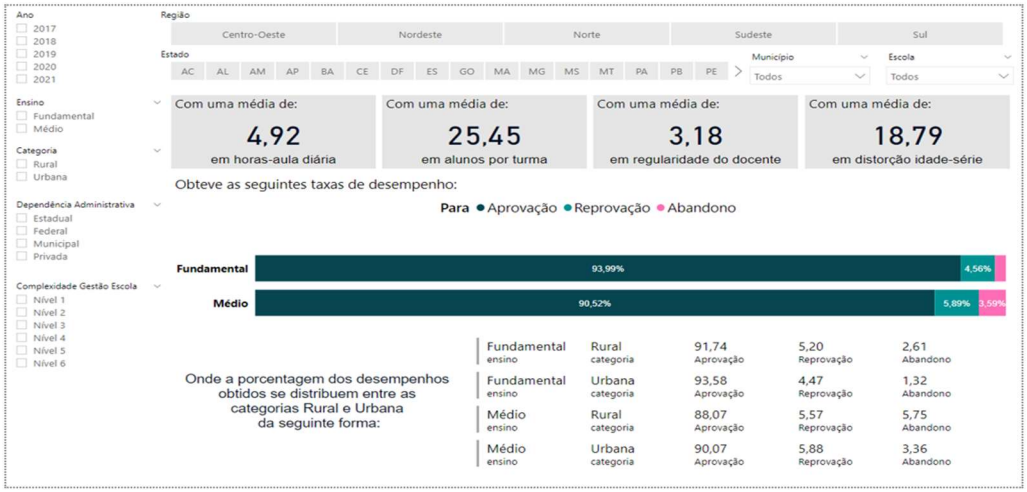

Figura 6. Análise do desempenho escolar sem segmentação
Fonte: Resultados originais da pesquisa

Conforme observado, durante o processo de criação do modelo de dados multidimensional, a estrutura existente entre as tabelas de dimensões e as tabelas fatos permite executar a ação de ajustar o cubo de dados para que se tenha as várias visões sobre os indicadores do desempenho escolar fornecidos pelo INEP (Nery, 2013). Por exemplo, a Figura 7 mostra o mesmo "dashboard", mas agora com a segmentação para um gestor educacional do estado de Goiás, que deseja analisar apenas os indicadores relacionados ao ano de 2017, que seja do ensino médio e com uma complexidade de gestão da escola de nível seis.

Figura 7. Análise do desempenho escolar com segmentação
Fonte: Resultados originais da pesquisa

Com a segmentação aplicada, a história a ser contada sobre os dados mudou o foco para o contexto em que foram aplicados os filtros. Nessa nova visão dos dados, não tem como, por exemplo, dizer algo sobre os dados do ensino fundamental, apenas sobre o ensino médio (já que essa era a vontade do gestor educacional ao aplicar o filtro). Com isso, uma história a ser contada sobre os dados poderia ser algo como: “No ano de 2017, na região Centro-Oeste, no estado de Goiás, com uma média de 3,06 na regularidade do docente, obteve-se uma taxa de aprovação de 89,79% para o ensino médio, onde as porcentagens dos desempenhos são distribuídas entre as categorias rural e urbana da seguinte forma: médio urbana 89,79% e médio rural inexistente, relacionados à aprovação e com a complexidade de gestão escola no nível seis (Knaflic, 2019).

Uma importante análise a ser feita sobre os dados está relacionada à evolução ao longo de um período (Nery, 2013). Conforme abordado, durante a criação do modelo multidimensional e na preparação dos dados, todos os arquivos do INEP possuem uma coluna referente ao ano para identificar o contexto temporal dos dados. Com isso foi possível criar uma estrutura visual para acompanhamento dos indicadores ao longo dos anos (Kimball e Ross, 2013). Em DATAVIZ os gráficos como o de linhas e barras são os mais indicados para esse tipo de análise e foram aplicados para fornecer uma análise das taxas de aprovação, reprovação e de abandono das escolas entre os anos de 2017 até 2021 (Sadiku et al., 2016). A Figura 8 mostra o resultado de uma visão sobre os dados levando em consideração a evolução temporal. Os dados apresentados têm as seguintes segmentações aplicadas: região Centro-Oeste, estado de Goiás, ensinos fundamental e médio, categoria urbana, dependência administrativa estadual, complexidade de gestão da escola de nível três.

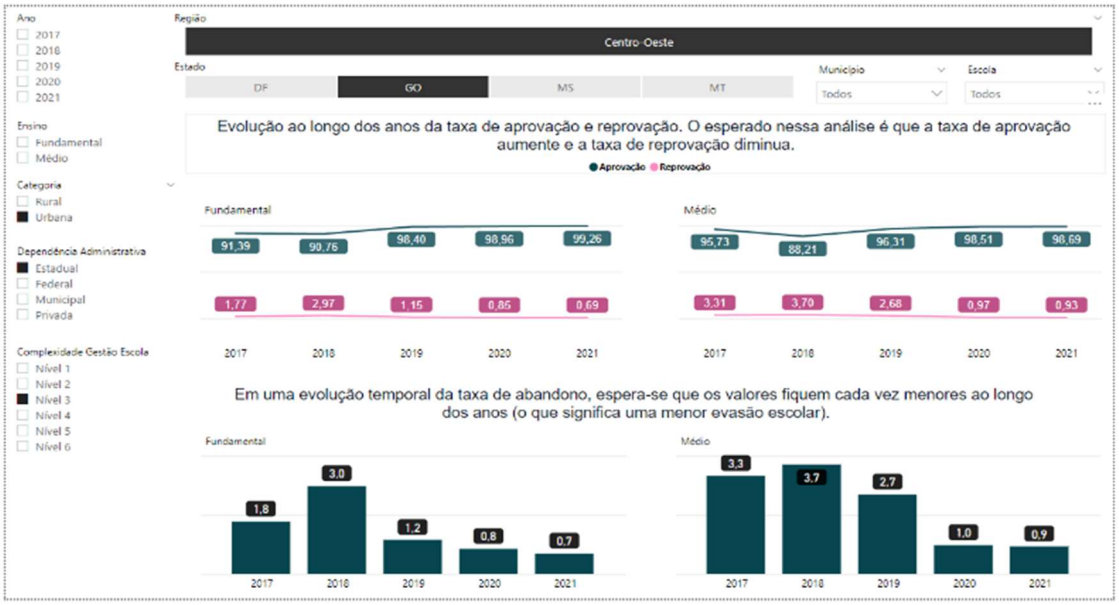

Figura 8. Análise temporal do desempenho escolar com segmentação
Fonte: Resultados originais da pesquisa

Ao analisar a segmentação apresentada, é possível notar uma certa estabilidade nas taxas de aprovação, reprovação e abandono no estado de Goiás, tanto para o ensino fundamental, quanto para o médio. Existem dois pontos que chamam a atenção. Para o ano de 2018 ambos os ensinos tiveram uma queda na taxa de aprovação e um aumento na taxa de abandono, já para o ano de 2021 (que é o mais recente nos dados) as taxas de abandono foram as menores no período, o que é algo positivo do ponto de vista da evasão escolar pois representa uma diminuição nos últimos três anos.

Seguindo a abordagem de análise temporal, foi elaborado um “dashboard” para acompanhamento da dependência administrativa (federal, municipal, estadual, privada) e para a categoria da escola (rural ou urbana). A Figura 9 mostra o resultado de uma visão sem segmentação, voltada à um gestor educacional nacional e contendo apenas os dados relacionados à taxa de aprovação.

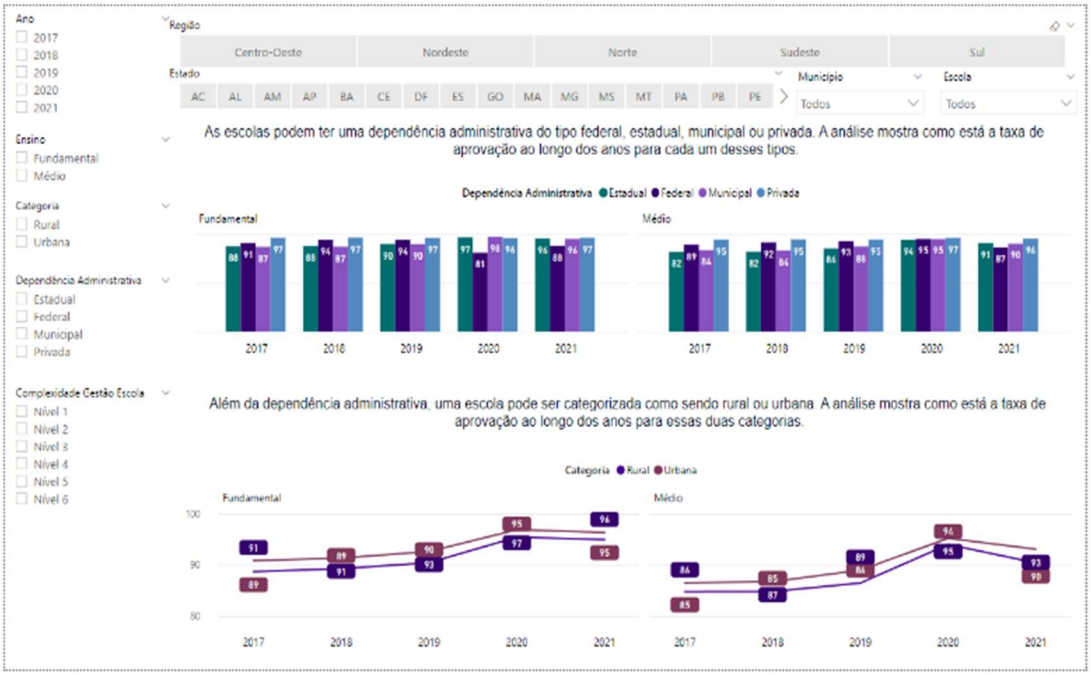

Figura 9. Análise temporal para dependência administrativa e categoria
Fonte: Resultados originais da pesquisa

É observado um certo desequilíbrio nas taxas de aprovação das escolas para todas as dependências administrativas ao longo de 2017 a 2021. Já nas categorias rural e urbana observa-se um crescimento positivo na taxa de aprovação para ambas, porém em 2021 a aprovação no ensino fundamental urbano teve uma pequena queda e o fundamental rural ficou estável. Já para o ensino médio urbano e rural, ambos tiveram uma queda no ano de 2021.

Seguindo a abordagem de DATAVIZ, pode ser que os dados apresentados estejam dificultando visualmente uma análise entre apenas dois tipos de dependências administrativas, como por exemplo uma análise entre as dependências administrativas

estadual e privada (Silva, 2019). Além disso, pode ser do desejo do gestor educacional ver apenas os dados da categoria urbana. Devido a isso, foi feita uma segmentação correspondente e foi apresentado o resultado na Figura 10, observe que agora existe uma segmentação para o estado de Goiás e com os filtros supracitados aplicados.

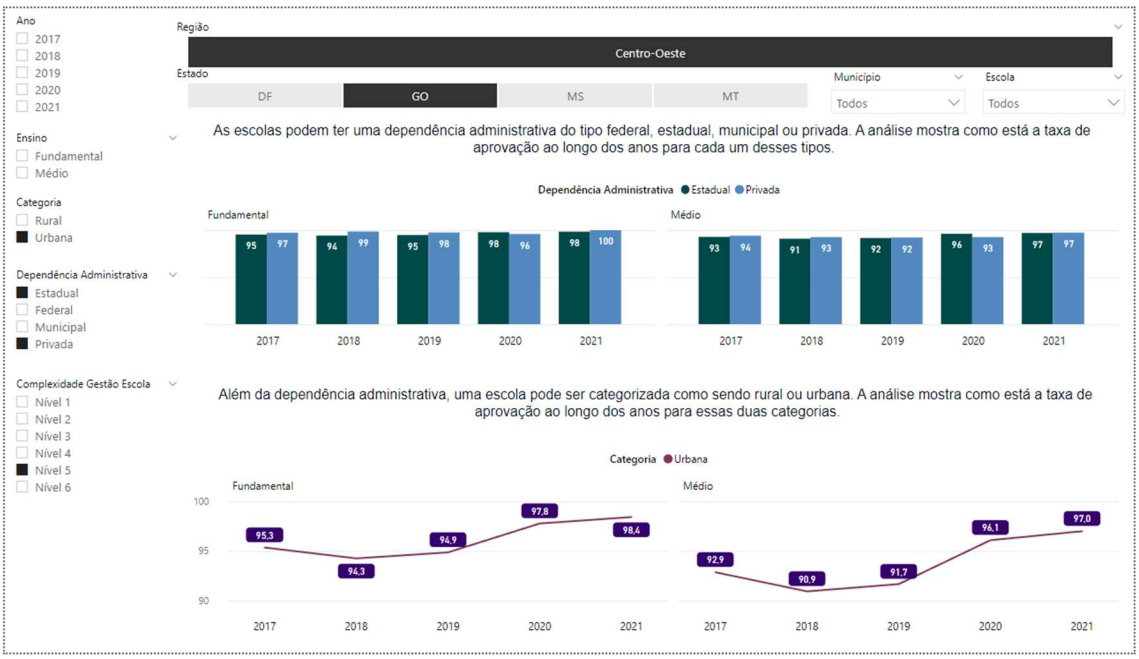

Figura 10. Análise temporal para dependência administrativa e categoria com segmentação
Fonte: Resultados originais da pesquisa

A segmentação aplicada permite ter uma melhor visualização nos gráficos de barras e de linhas existentes na tela (Silva, 2019). Agora percebe-se com mais clareza que a dependência administrativa privada tem uma maior taxa de aprovação na maioria dos casos e que apenas em 2020 a estadual ficou à frente. Por outro lado, uma análise apenas sobre a evolução urbana ficou melhor de ser observada e percebe-se que para essa segmentação a taxa de aprovação teve um crescimento positivo ao longo dos últimos três anos (Silva, 2019).

O arquivo IGR fornecido pelo INEP (Tabela 3) contém a classificação das escolas em níveis que vão do um ao seis, onde quanto maior for o nível, maior será a complexidade de gestão da escola (INEP, 2022). Durante as fases anteriores, foi identificado que este arquivo é o único dentre os demais que não possui um valor quantitativo para métrica pois é uma variável categórica, todavia, aplicando a técnica apropriada foi criada a dimensão D_Nivel que é usada como filtro nas segmentações (Nery, 2013). Na fase de negócio foi levantado que os gestores educacionais queriam ter uma visão sobre a quantidade de escolas em cada um desses níveis com a respectiva taxa de aprovação, a Figura 11 mostra o “dashboard” criado para atender essa finalidade. Para Silva (2019) a forma de um gráfico depende do que se quer transmitir. Fez-se então o uso do gráfico de barras, que é o indicado

dentro de uma abordagem de DATAVIZ para a apresentação de variáveis categóricas (Knaflic, 2019).

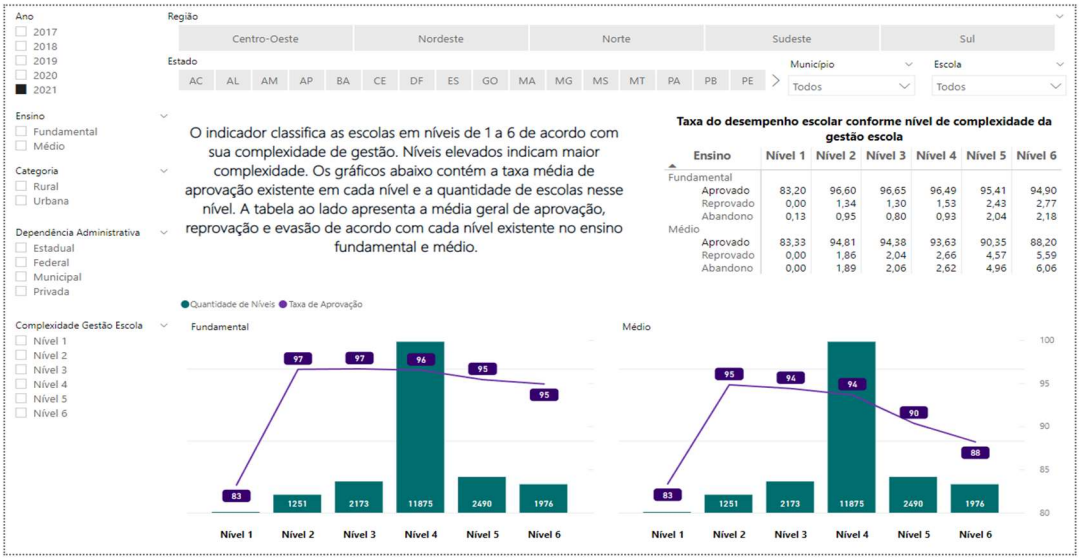

Figura 11. Análise do nível de complexidade da gestão escola com segmentação
Fonte: Resultados originais da pesquisa

Na imagem é possível observar que houve uma segmentação para o ano de 2021. Os ensinos fundamental e médio têm a mesma quantidade porque estão sendo trabalhadas apenas escolas que contém ambos os ensinos. Existem seis escolas classificadas no nível um da complexidade de gestão escola e com uma média de taxa de aprovação de 83 para ambos os ensinos. Existem 1251 escolas no nível dois da complexidade de gestão escola onde o ensino fundamental tem uma taxa média de aprovação de 97 e o ensino médio uma taxa média de aprovação de 95. Existem 2173 escolas no nível três da complexidade de gestão escola onde o ensino fundamental tem uma taxa média de aprovação de 97 e o ensino médio tem uma taxa média de aprovação de 94. E assim segue a análise até o nível seis de complexidade da gestão escola. No resultado observa-se que conforme vai aumentando o nível de complexidade vai diminuindo a taxa média de aprovação, com exceção do nível um que segundo o INEP tem a complexidade mais baixa e está com uma taxa de aprovação bem abaixo dos demais, porém deve ser analisado mais detalhadamente para identificar o fenômeno (talvez seja devido à quantidade das escolas nesse nível serem extremamente poucas. Neste caso deve-se investigar melhor o acontecimento, podendo evoluir até mesmo para uma análise sobre os métodos de avaliação aplicados pelo INEP).

“Deployment”

Caso todas as etapas anteriores tenham sido elaboradas da maneira correta, essa será a última etapa do CRISP-DM (Huber et al., 2019). É aqui que os modelos desenvolvidos

deverão ser colocados em produção, visando a disponibilização dos recursos a serem acessados pelas partes interessadas em realizar as análises sobre os dados, no caso desse trabalho os gestores educacionais (Colpani, 2018). A forma como essa etapa é feita varia muito, pois depende do tipo de projeto, do modelo aplicado e dos recursos trabalhados (Martínez et al., 2019). Fatores como onde serão alocados os recursos influenciam muito, por exemplo, se irá ficar em um servidor próprio na empresa (“on premisse”) ou se irá ficar em um servidor na nuvem, etc. No caso desse trabalho, como foi desenvolvido em “Power” BI, existem duas possibilidades. A primeira é a geração e disponibilização do arquivo que contém todos os recursos construídos, tais como os dados e componentes visuais. A segunda possibilidade, que é a mais indicada, seria colocar em um servidor do “Power” BI e disponibilizar o acesso aos gestores educacionais (Lachev, 2022).

Considerações Finais

Toda contribuição para prover uma melhoria na educação é muito importante. A disponibilização dos indicadores de desempenho escolar é um ponto de partida para esse intento. Seguindo essa linha, a criação de uma estrutura voltada para a análise desses indicadores em que sejam trabalhados conceitos, metodologias e técnicas existentes na ciência de dados, poderão agregar e muito no dia-a-dia daqueles que estão à frente de uma gestão educacional.

A partir desta pesquisa e com o uso de uma metodologia apropriada de mineração de dados para conduzir a aplicação dos conceitos de BI, onde é possível trabalhar os dados fornecidos e ter uma estrutura que permita múltiplas visões a serem apresentadas através de um resultado visual bem elaborado, com a concentração de dados outrora dispersos, é observável que o acompanhamento dos indicadores de desempenho poderão ser utilizados de uma forma mais oportuna dentro de um contexto voltado à tomada de decisão, e que poderá contribuir com o processo de melhoria do ensino brasileiro. Uma possível aplicação seria em um cenário onde um gestor educacional acabou de assumir a escola e deseja obter múltiplas visões sobre os indicadores do INEP e assim poder analisar como está a evolução daquela unidade nos últimos anos. Uma outra possibilidade, como os indicadores fornecidos pelo INEP estavam em arquivos separados, por exemplo a taxa de desempenho escolar, quantidade média de alunos por turma e a taxa de distorção idade-série, agora estes poderão ser analisados em conjunto e verificar se a aprovação, reprovação ou até mesmo a evasão escolar tem sofrido impactos por tais indicadores e assim tomarem decisões sobre a possibilidade em se criar novas turmas para diminuir uma possível lotação existente ou até mesmo realizar um trabalho voltado para a melhoria nos indicadores de desempenho impactados por questões relacionados à faixa-etária dos alunos.

Agradecimentos

Agradeço à Deus pela superabundante graça e pela minha existência, à Jesus Cristo pela expiação e nova aliança, ao Espírito Santo pelo consolo e fortalecimento da fé. Agradeço também aos amigos, familiares, orientação e aos que compartilharam o conhecimento.

Referências

Chawla, G.; Bamal, S.; Khatana, R. 2018. Big data analytics for data visualization: review of techniques. International Journal of Computer Applications 182(21): 37-40.

Colpani, R. 2018. Mineração de dados educacionais: um estudo da evasão no ensino médio com base nos indicadores do censo escolar. Informática na educação teoria e pratica, Porto Alegre, RS, Brasil. Disponível em: https://www.seer.ufrgs.br/index.php/InfEducTeoriaPratica/article/view/87880. Acesso em: 9 jan. 2023.

Diamond, M.; Mattia, A. 2017. Data visualization: An exploratory study into the software tools used by businesses. Journal of Instructional Pedagogies 17: 1-7.

Fonseca, S. O.; Namen, A. A. 2016. Mineração em base de dados do INEP: Uma análise exploratória para nortear melhorias no sistema educacional brasileiro. Educação em Revista 32: 133-157.

Huber, S.; Wiemer, H.; Schneider D.; Ihlenfeldt, S. 2019. Data mining methodology for engineering applications: a holistic extension to the CRISP-DM model. Procedia CIRP 79: 403-408.

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira [INEP]. 2022. Indicadores de desempenho escolar. Disponível em: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/indicadores-educacionais. Acesso em: 9 jan. 2023.

Kimball, R.; Ross, M. 2013. The data warehouse toolkit: the definitive guide to dimensional modeling. 3ed. Wiley, Hoboken, NJ, USA.

Knaflic, C. N. 2019. Storytelling com dados: um guia sobre visualização de dados para profissionais de negócios. 2ed. Alta Books, Rio de Janeiro, RJ, Brasil.

Lachev, T. 2022. Applied Microsoft Power BI: Bring your data to life. 7ed. Prologika Press, USA.

Laureano, R. M. S.; Caetano, N.; Cortez, Paulo. 2014. Previsão de tempos de internamento num hospital português: aplicação da metodologia CRISP-DM. Risti 13 :83-98

Martínez-Plumed, F.; Contreras-Ochando, L.; Ferri, C.; Hernández-Orallo, J.; Flach, P.; Kull, M.; Lachiche, N.; Ramirez-Quintana, M. J. 2019. CRISP-DM twenty years later: from data mining processes to data science trajectories. IEEE Transactions on Knowledge and Data Engineering 33: 3048-3061.

Nascimento, L. S.; Cruz Jr., R. G.; Fagundes, G. A. A. 2018. A mineração de dados educacionais: um estudo sobre indicadores da educação em bases de dados do INEP. Renote, Porto Alegre, RS, Brasil. Disponível em: https://seer.ufrgs.br/index.php/renote/article/view/85989. Acesso em: 9 jan. 2023.

Nery, F. R. 2013. Tecnologia e projeto de data Warehouse: uma visão multidimensional. 6ed. Érica Saraiva, São Paulo, SP, Brasil.

Plotnikova, V.; Dumas, M.; Milani, F. 2022. Applying the CRISP-DM data mining process in the financial services industry: Elicitation of adaptation requirements. Data Knowledge Engineering 139: 102013-102030.

Provost, F.; Fawcett, T. 2016. Data Science para Negócios. 1ed. Alta Books, Rio de Janeiro, RJ, Brasil.

Sadiku, M.; Shadare, A. E.; Musa, S. M.; Akujuobi, C. M.; Perry, R. 2016. Data visualization. International journal of engineering research and advanced technology 2(12): 11-16.

Silva, F. C. C. 2019. Visualização de dados: passado, presente e futuro. Liinc em revista 15(2): 205-223.