UNIVERSIDADE
FEDERAL RURAL
DE PERNAMBUCO

Abílio Nogueira Barros

Elaboração do conjunto de dados agregados do censo da educação básica

Recife

2022

Abílio Nogueira Barros

Elaboração do conjunto de dados agregados do censo da educação básica

Artigo apresentado ao Curso de Bacharelado em Ciências da Computação da Universidade Federal Rural de Pernambuco, como requisito parcial para obtenção do título de Bacharel em Ciências da Computação.

Universidade Federal Rural de Pernambuco – UFRPE

Departamento de Computação

Curso de Bacharelado em Ciências da Computação

Orientador: Rafael Ferreira Mello

Recife

2022

Agradecimentos

Agradeço primeiramente a Deus e logo após minha filha, que na figura principalmente dos meus pais,Aparecida e Erigerson, foram de vital inspiração e suporte durante toda a jornada da graduação, dando sentido a todo o tempo e energia investidos para chegar até o presente momento. Agradeço também ao meu orientador ,Rafael Mello,que me guiou durante todo o percurso da realização desse e de tantos outros projetos que moldaram o conhecimento e experiência necessária para a realização desse projeto. É necessário mencionar, mesmo que não nominalmente, outros docentes do departamento da computação que passaram, além do conhecimento, sua experiência e dividiram sua vivência acadêmica, como também todo o suporte fornecido pela pessoa responsável do secretariado do curso, que sempre esteve o mais próximo possível para auxiliar nas questões burocráticas da vivência acadêmica. É preciso citar os agradecimentos a toda a estrutura da Universidade Federal Rural de Pernambuco que em sua maioria faz o melhor possível, dentre as condições, para promover o ensino público de qualidade. E por fim, mas jamais menos importante, meus amigos, tanto os vindos da escola quanto o da graduação, que foram um pilar tal qual importante,para toda essa jornada e todas as outras que venham a seguir, sempre dividindo todos os momentos dessa jornada acadêmica.

Resumo

O Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) disponibiliza o Censo da Educação Básica, o maior levantamento anual de dados sobre a educação brasileira. Os dados são disponibilizados anualmente e com cerca de 370 colunas e pouco mais de 230 mil registros por ano. Este trabalho apresenta o processo que foi utilizado para criar um conjunto de dados que unificasse os anos de 2010-2021 e o disponibilizasse de forma a garantir boas práticas de disponibilização de dados na web. Foi gerado um conjunto de dados abrangendo todos os anos mencionados, posteriormente dividido em subconjuntos dada a natureza dos dados apresentados.

Palavras-chave: educação, dados, educação básica.

Abstract

The Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) makes available the Census of Basic Education, the largest survey of data on Brazilian education every year. The data are made available annually and with about 370 columns and just over 230 thousand records per year. This work presents the process that was used to create a dataset that would unify the years 2010-2021 and make it available in order to ensure good practices for making data available on the web. A dataset was generated by creating a sub-division of the provided data and aligning its data dictionary to reflect the current context of the data produced.

Keywords: education, data, basic education

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
TSE	Tribunal Superior Eleitoral
ENEM	Exame Nacional do Ensino Médio
SAEB	Sistema Nacional de Avaliação da Educação Básica
ENADE	Exame Nacional de Desempenho dos Estudantes
ETL	*Extract, transform, load*
CSV	*Comma-separated values*
PNE	Plano Nacional de Educação
API	*Application Programming Interface*

Sumário

1 Introdução

Atualmente existem muitas variáveis que podem impactar a tomada de decisão na sociedade como visto em (LAI; SCHILDKAMP, 2013), porém para a utilização dessas variáveis se faz necessários dados para que possam ser estudadas e levadas em consideração para tomada de decisão. Muitos trabalhos têm utilizado dados para dar suporte a essa tomada de decisão. Na área de educação não é diferente, muitas são as possibilidades da utilização de dados para ajudar professores e gestores a decidir quais passos seguir na construção de um contexto educacional favorável ao desenvolvimento do aluno(JAMES; MILENKIEWICZ; BUCKNAM, 2008).

No Brasil, o INEP (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira)[1] sintetiza anualmente dados sobre a educação básica e superior . São várias fontes de dados que estão disponibilizadas de forma pública e que podem ajudar gestores em suas ações diárias. Contudo, apenas o acesso aos dados não é o suficiente para gerar informação necessária para auxiliar os gestores (FRENEDA et al., 2020). Além disso, devido a complexidade do contexto educacional no Brasil, em alguns anos o INEP mudará variáveis para melhor apresentar os resultados, mas isso também gera uma incompatibilidade dos dados. Por isso, em alguns momentos existem diferentes plataformas consumindo os mesmos dados, mas com resultados diferentes (BARRETO; FREITAS, 2020).

Sendo assim, o censo educacional é um instrumento fundamental para disponibilizar dados sobre as instituições públicas e privadas do país(DINIZ, 1999). Atualmente a tabela de dados original é disponibilizada com mais de trezentos atributos, tornando assim inviável unir todos os anos sem o devido processamento e fornecimento de forma organizada dos dados por temática e relação, dado que todas as variáveis são informadas no mesmo arquivo.

Dado a quantidade de informações unitárias , a nível de instituição escolar, disponibilizadas para cada instituição escolar, pouco menos de quatrocentos atributos, a separação desses atributos por temática pode trazer uma maior facilidade na hora da utilização desses dados, visto que hoje são todos disponibilizados no mesmo arquivo e no momento que forem combinados numa série histórica acabaria gerando grande volumes de dados a depender do poder de processamento disponível para o usuário durante a utilização dos dados.

Ainda não é possível se encontrar esses dados todos condensados na versão atual do censo educacional disponibilizado pelo INEP, afinal são disponibilizados anu-

[1] https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados

almente, e sua junção demanda tratamento dos dados para que possam ser ajustados para um mesmo registro contendo todos os anos alvos, validando que os utilizados neste trabalho serão de 2010-2021, sendo assim é necessária uma forma de unir esses dados e disponibilizar de uma forma acessível. Para as versões anteriores do censo existem alguns fornecedores desses dados, um deles é o Laboratório de Dados Educacionais [2], onde fornecem alguns dados de forma processada os dados censitários de 2007-2020, porém não todas as variáveis disponibilizadas pelo INEP.Podendo também ser destacado como dificuldade o fato que para carregarmos apenas dados de matrículas das escolas, seria necessário realizar a carga de todos os arquivos dos anos de interesse e a posterior selecionar as colunas de interesse, uma vez que na atual forma os dados são disponibilizados em um mesmo arquivo.

Se faz necessário além de disponibilizar esses dados de forma acessível e de simples utilização, é preciso seguir boas práticas para a publicação desses dados na internet. Fornecendo assim *metadados* para que qualquer indivíduo ou grupo que deseje utilizar os dados saibam que tipos de dados e qual a descrição de cada coluna dos registros ali disponíveis.

A estrutura desse documento se apresenta da seguinte forma: Na seção 2 são apresentados trabalhos correlatos tanto de utilização desses dados educacionais; a seção 3 descreve como foi a coleta e processamento dos dados retirados do INEP; já para a seção 4 temos a forma da disponibilização e exemplos de utilização desses dados; por último temos na seção 5 as considerações finais debatendo os resultados desse trabalho e trabalhos futuros.

[2] https://dadoseducacionais.c3sl.ufpr.br//

2 Trabalhos Relacionados

No campo de construção de bases de dados de bases públicas é importante destacar o caso demonstrado em (VASCONCELOS et al., 2021), onde foram recolhidos e padronizados dados diretamente do portal do TSE. Após etapas de aquisição e processamento dos dados brutos, foram adicionados novos dados geográficos com o objetivo de fornecer mais usabilidade aos utilizadores das bases.

Transportando a temática para a área da saúde temos em (GONÇALVES et al., 2021) a coleta, tratamento e disponibilização de base de dados sobre a vacinação registrados na plataforma **OpenDataSUS**. O processo se baseou sobre características de processos de ciência de dados, realizando seu desenvolvimento de forma interativa até atingir o objetivo de disponibilização do conjunto de dados e dicionário de dados.

Migrando para a área educacional, já existem alguns trabalhos nessa linha de utilização de dados educacionais para tomada de decisão, principalmente na utilização das bases de dados fornecidas pelo INEP. Em (CONTE, 2019) Utilizou as bases de perfil socioeconômico e resultados do Exame Nacional do Ensino Médio (ENEM), podendo assim realizar procedimentos de limpeza e organização dos dados a fim de fornecer conhecimento sobre a base do exame na edição do ano de 2015. Assim como destacado pelo autor, faz-se de extrema importância associar esse conhecimento adquirido com as outras bases fornecidas como Prova Brasil, ENADE e SAEB.

Em (FILHO; ISOTANI; PENTEADO, 2021) foram utilizadas notas de disciplinas escolares para a predição dos valores dos resultados dos alunos no ENEM (Exame Nacional do Ensino Médio). Foi definido uma quantidade amostral do grupo completo dos alunos de uma determinada escola, foi realizado o levantamento de sua pontuação dentro do âmbito escolar e posteriormente cruzado com sua pontuação do ENEM informada pelos alunos que participaram do estudo. Trazendo assim, mesmo que de maneira inicial, o impacto das notas dos alunos ainda na unidade escolar possa ser fator relevante para os resultados na área de ciência humanas do ENEM.

Foram realizados também trabalhos focando especificamente numa rede educacional. Como descrito em (PINTO; JÚNIOR; COSTA, 2019), foram trabalhados resultados das rede municipal da cidade de Maceió com o intuito de poder extrair as melhores *features* dos alunos da localidade gerando esse conhecimento específico sobre os alunos daquela localidade. Foi apresentado em (WANDERLEY et al., 2021), o uso de alguns indicadores selecionados pelo autor para o acompanhamento das escolas do Acre. Mapeando as fontes de dados do INEP para criar dez painéis utilizando a ferramenta *Power BI*.

Diante disso, o objetivo desse trabalho é fornecer o conjunto de dados que aborda o panorama geral dos dados educacionais na versão mais atual fornecida pelo INEP, dos anos de 2010 a 2021 onde possam ser filtrados e agregados tanto por unidade escolar quanto por município, e ainda poder acompanhar a mudança dessas características através dos anos. Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Donec odio elit, dictum in, hendrerit sit amet, egestas sed, leo. Praesent feugiat sapien aliquet odio. Integer vitae justo. Aliquam vestibulum fringilla lorem. Sed neque lectus, consectetuer at, consectetuer sed, eleifend ac, lectus. Nulla facilisi. Pellentesque eget lectus. Proin eu metus. Sed porttitor. In hac habitasse platea dictumst. Suspendisse eu lectus. Ut mi mi, lacinia sit amet, placerat et, mollis vitae, dui. Sed ante tellus, tristique ut, iaculis eu, malesuada ac, dui. Mauris nibh leo, facilisis non, adipiscing quis, ultrices a, dui.

3 Aquisição e Processamento dos dados

O processo de desenvolvimento desse conjunto de dados foi desenvolvido tendo como base o processo de ETL (Extract, Transform, Load) assim como definido em (FERREIRA et al., 2010) é um processo de extrair os dados de suas fontes, de preferências originais, aplicarem os processos necessários a fim de transformar aquela base da forma mais acessível e direta para a utilização, e o *Load* vem do carregamento dentro a um local de armazenamento mais definitivo, onde no escopo deste projeto será utilizado um conjunto de arquivos no formato CSV (Comma-separated values).

Todo o processamento executado foi projetado tendo como base o objetivo final de consolidar aquela série temporal de 2010 a 2021 em um conjunto de dados que pudesse ser incorporado para uso já em sua forma de saída do processo. Para o correto uso dos dados, o fluxo aplicado foi desenhado após a leitura e entendimento dos dicionários de dados oferecidos pelo disponibilizado oficial dos dados. Tendo assim o fluxo de dados ilustrado na figura 1.

3.0.1 Extração dos dados

Os dados originais deste projeto foram inteiramente retirados do site do INEP. Esses dados são fornecidos no formato compactado a fim de facilitar o descarregamento para a máquina local. Os dados originalmente são fornecidos anualmente, são fornecidos anualmente um arquivo no formato CSV contendo os dados e um dicionário de dados no formato de planilhas do excel, além de outros *metadados* como: os questionários aplicados às instituições; um manual do usuário indicando informações para uso em determinados *softwares*. As bases originais continham 370 colunas e variando entre 200876 a 224229 registros anuais, totalizando 2.792.984 registros.

A coleta de dados foi feita por meio de um *script* na linguagem Python, onde posteriormente foi salva em diretórios da máquina onde foi desenvolvido o projeto. Logo após foi checado manualmente se todos os anos alvos estavam de fato contidos no diretório em questão. Para os passos seguintes o mesmo *script* iria descompactar os dados para que pudesse ser realizada a leitura manual dos dicionário de dados disponibilizados para que pudesse ser elaborada a etapa de padronização dos dados.

3.0.2 Transformação dos dados

Após a leitura dos dicionários de dados foi possível elencar quais alterações eram necessárias serem aplicadas com o objetivo de unir todos os anos que foram elencados como objetivo do projeto. Foi observado que grande parte das colunas in-

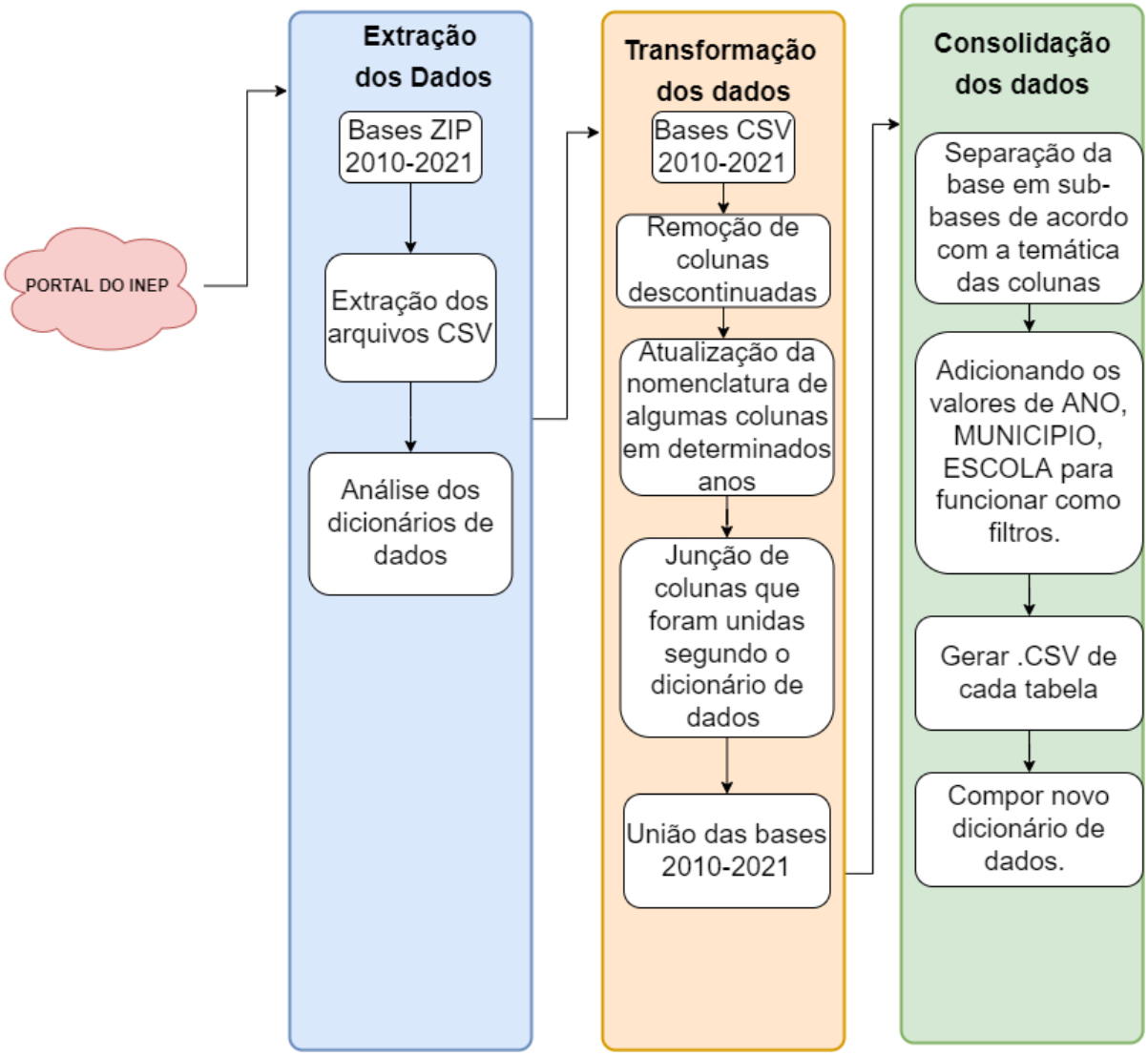

Figura 1 – Fluxo de processamento dos dados utilizado.

formadas já foram padronizadas, visto que já existiram atualizações realizadas pelo próprio INEP nas versões através dos anos, entretanto existiam colunas que ainda impediriam essa junção e algumas outras que foram descontinuadas em relação aos anos mais recentes.

A transformação de dados foi desenvolvida em algumas etapas que são relatadas nas próximas subseções.

3.0.2.1 Remoção de colunas descontinuadas

Foram removidas as colunas que foram descontinuadas a partir de 2019, pois os dados mais atuais não iram mais refleti-las e assim não devem trazer informações tão atuais para o acompanhamento da estrutura da rede educacional.Do total das 370 variáveis disponibilizadas, 36 foram removidas. Restando assim 334 para o processa-

mento das etapas seguintes. Uma lista com as colunas removidas foi disponibilizada[1].

3.0.2.2 Atualização da nomenclatura das colunas

Seguindo para o processamento das colunas remanescentes certas alterações precisaram ser aplicadas para que fosse possível uma adequação de nomenclatura e variáveis para a faixa temporal aplicada. Tal necessidade foi apontada pela leitura dos dicionários de dados que foram disponibilizados junto aos arquivos, podendo assim comparar as alterações pelos campos de *Observações* e *Coleta por ano*.Para o processamento das colunas temos duas etapas:

3.0.2.3 Junção das colunas

Algumas colunas tiveram seus nomes alterados depois de 2018, com isso foi necessário padronizar essas colunas para os nomes mais atuais. Colunas essas que não tiveram seus valores alterados, apenas a nomenclatura, esse processo foi realizado seguindo o padrão indicado pelo dicionário de dados.

Outras colunas tiveram que ser unidas a partir do ano de 2019, sendo assim será utilizado a estratégia do **OU**, pois essas colunas são especificações de uma mesma resposta, sendo assim basta que apenas uma das colunas tenham seu resultado atendido que será aplicado o valor de Verdadeiro para a nova coluna, caso não tenham nenhum valor positivo o valor falso será atribuído a coluna. A quantidade de registros total permaneceu inalterada.

3.0.3 Consolidação dos dados

Os processos que foram citados anteriormente foram aplicados individualmente em cada ano, mantendo a quantidade de registros originais. Podemos acompanhar na tabela 2 a quantidade de registros e colunas a cada ano.

Somando a quantidade de valores temos o resultado de 2.792.984, mostrando assim que tal quantidade foi inalterada pelo processo.

Dado a quantidade de colunas e buscando manter o padrão antigo de disponibilização do censo educacional, onde os dados eram informados em uma divisão de tabelas em que cada tabela fornece dados sobre uma temática específica como turma;professores;escola. Sendo assim e com o objetivo de facilitar o uso de dados para consultas específicas, como por exemplo **Qual a quantidade de matrículas de uma determinada escola ao decorrer dos anos ?**, para conseguir responder tal pergunta não será necessário carregar todas as colunas.

[1] https://docs.google.com/spreadsheets/d/1qjZA6YAuUvWum5MpibKo0Hp$_{vpVeyCavs5ZGy59}$ $-_{m}$ $Q/edit?usp = sharing$

Ano	Quantidade de Colunas	Quantidade de Linhas
2010	328	200876
2011	328	242147
2012	328	242136
2013	328	242680
2014	328	242929
2015	328	237879
2016	328	237506
2017	328	236481
2018	328	236460
2019	328	228521
2020	328	224229
2021	328	221140

Tabela 1 – Quantidade de registros pós processamento.

Nome da Tabela	Quantidade de Colunas	Tamanho do arquivo(MB)
TB_GEOGRAFICA	14	329.4
TB_TURMA	19	219.3
TB_DOCENTE	19	220.9
TB_MATRICULA	43	485.1
TB_ESCOLA_ADMINISTRATIVO	31	431.8
TB_ESCOLA_ESTRUTURA	116	854.6
TB_ESCOLA_FUNCIONARIO	33	194.9
TB_ESCOLA_PEDAGOGICO	69	481.8

Tabela 2 – Divisão das bases

A divisão da base buscou alinhar o máximo possível as variáveis em conjuntos de significado de seus dados. Como a base de dados original é orientada a escola, foi preciso subdividir as informações diretamente ligada à entidade escolar.

Nome da Tabela	**Descrição**
TB_GEOGRAFICA	Dados geográficos onde a instituição escolar está alocada.
TB_TURMA	Informações sobre as turmas de uma determinada instituição escolar.
TB_DOCENTE	Dados quantitativos sobre os docentes pertencentes a instituição escolar.
TB_MATRICULA	Matrículas realizadas na instituição escolar.
TB_ESCOLA_ADMINISTRATIVO	Dados sobre a instituição escolar.
TB_ESCOLA_ESTRUTURA	Dados sobre a estrutura da instituição escolar.
TB_ESCOLA_FUNCIONARIO	Dados sobre os funcionários não docentes da instituição escolar.
TB_ESCOLA_PEDAGOGICO	Dados sobre a estrutura pedagógica da instituição escolar.

Figura 2 – Conteúdo de cada tabela

3.0.4 Novo dicionário de dados

O INEP disponibiliza o dicionário de dados anualmente junto a sua respectiva base de dados, dicionário esse totalmente necessário para poder acompanhar os da-

dos representado nas centenas de campos ali informados, além de descrever qual tipo, e em alguns casos, as categorias ali representada por valores inteiros.

Ao realizar a concatenação dos dados anuais e uma nova divisão da base censitária se faz necessário gerar uma nova base de dados que possa refletir o atual estado do conjunto de dados. Esse novo dicionário é baseado no fornecido pelo INEP e considerando a divisão atual das bases para que seja possível realizar o acompanhamento e junção , quando necessário pelo utilizador dos dados. Dicionário esse que apresentará a faixa temporal atual da base e será fornecido junto ao conjunto de dados.

3.0.5 Boas práticas adotadas

Para uma boa aplicação desses dados são necessário além da disponibilização desses dados na web um suporte para sua utilização, existem um conjunto de boas práticas [2] a serem adotadas ao disponibilizar conjunto de dados na internet. Assim sendo, a estrutura desse projeto listou se baseou nas seguintes boas práticas a fim de garantir que seja possível a utilização de tal conjunto de dados por qualquer um que deseje extrair informações do mesmo. Podemos observar na tabela a seguir quais boas práticas foram implementadas no projeto. Ao alcançar as metas apontadas na

Código	Definição
BP1	FORNECER METADADOS
BP2	FORNECER METADADOS DESCRITIVOS
BP3	FORNECER METADADOS ESTRUTURAIS
BP5	FORNECER INFORMAÇÕES DE PROVENIÊNCIA DOS DADOS
BP7	FORNECER INDICADOR DE VERSÃO

Tabela 3 – Boas práticas atingidas pela base

tabela 3 conseguimos fornecer uma base dados que seja possível ser rastreável tanto quanto a seus dados de origem (BP5), quanto qual versão está sendo utilizada e qual época de publicação (BP7), além da disponibilização dos metadados atualizados para uma coerente utilização dos dados (BP1,BP2,BP3).

[2] https://ceweb.br/media/docs/publicacoes/1/fundamentos-publicacao-dados-web.pdf

4 Disponibilização e Utilização

4.0.1 Disponibilização dos dados

Tanto o *dataset* **Conecta PNE** quanto seus metadados necessário para a utilização do conjunto de dados estão sendo disponibilizados no seguinte link da plataforma ZENODO[1] <https://zenodo.org/record/6666613#.YrioP3bMLnY>. Os arquivos estão dispostos sobre a licença CC-BY-4.0 onde, esta licença permite que outros baixem, ajustem e desenvolvam esse trabalho, mesmo que seja utilizado para fins comerciais, contando que esse projeto seja referenciado como base.

4.0.2 Possíveis Cenários para utilização

Por concentrar dados escolares de instituições de todo o Brasil, a aplicação desse conjunto de dados pode ser voltada a todo tipo de exploração do panorama de mais de uma década da educação básica brasileira. O objetivo com a disponibilização desses dados é prover uma forma que além de ,gestores de instituições e secretarias estaduais e municipais, que grupo de pesquisas, alunos que possam conduzir pesquisas na área e quaisquer um que deseje saber mais como se encontra e quais avanços e retrocessos foram registrado na educação de seu entorno.

É possível também destacar os cenários de utilização como acompanhamento de metas do PNE (Plano Nacional da Educação) onde alguns indicadores podem ser calculados por atributos contidos na base. É possível também fazer a utilização de algoritmos de *machine learning* aplicado a variáveis de série histórica e assim realizar a estimativa de futuros cenários para aquela instituição ou rede de ensino.

Também se faz possível a utilização para o monitoramento de melhorias determinada entidade escolar ou até mesmo de uma rede escolar como um todo, podemos verificar atributos como a quantidade de docentes com determinado grau escolar ou até mesmo número de salas climatizadas ou materiais de informática para o uso no âmbito escolar, possibilidades essas contempladas no conjunto de dados.

4.0.3 Exemplos de Aplicação

São inúmeras possibilidades de aplicação desses dados, como em (BALBINOT; HAUBERT, 2017) a aplicação desses microdados censitários foi utilizada para o acompanhamento de indicadores sobre educação especial no Rio de Janeiro, já em (BALBINOT; HAUBERT, 2015) foi realizada uma análise temporal das matrículas do estado

[1] https://zenodo.org/

do Paraná. São alguns exemplos sobre trabalhos já realizados com base nesses dados educacionais. Por se tratar de dados de fonte oficial, que é o INEP, pode ser também utilizado para acompanhamento desde uma instituição escolar em específico, pois se trata do menor nível de especificação dos dados, como ser agrupados até o nível estadual. Dados esses que também podem contribuir para o acompanhamento das metas do PNE (Plano Nacional da Educação)[2], podemos citar que as dentre as 3 primeiras metas cinco de seus 6 indicadores apresentados são calculados com a quantidade de matrículas realizadas a nível municipal, estadual e federal.[3] [4] [5]

Podemos também levantar nos próprios dados algumas visualizações sobre a situação escolar de um determinado município, buscando entender qual a situação refletida no tempo levantado pela base. Na figura 3 podemos ver o decréscimo da

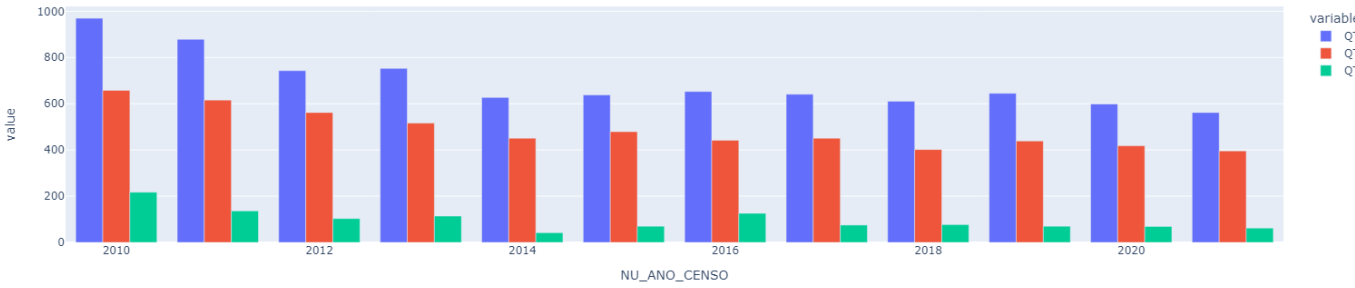

Figura 3 – Quantidade de Professores por ano e grau do Município de Moreno-PE

quantidade de professores do município de Moreno-PE, os professores das três etapas de ensino (Azul representando ensino básico, Vermelho o fundamental e verde o ensino médio), é de fácil constatação que naquela localidade a maior quantidade de professores disponíveis é do ensino básico e podemos buscar observar a quantidade de matrículas se estabelece esse mesmo tipo de padrão.

Seguindo na utilização da base, podemos acompanhar na figura 4 que, no mesmo exemplo de cores e legendas Azul representando matrículas do ensino básico, vermelho matrículas do ensino fundamental e verde matrículas do ensino médio), a quantidade majoritária de matrículas é pertencente a faixa de educação básica e que, igualmente a quantidade de professores da rede de ensino, a quantidade desse tipo de faixa educacional vem caindo porém ainda se mantendo em maior número em relação aos demais.

[2] https://pne.mec.gov.br/
[3] http://simec.mec.gov.br/pde/pne/notas_tecnicas/Nota_Tecnica_Meta_1_ciclo_1.pdf
[4] http://simec.mec.gov.br/pde/pne/notas_tecnicas/Nota_Tecnica_Meta_2_ciclo_1.pdf
[5] http://simec.mec.gov.br/pde/pne/notas_tecnicas/Nota_Tecnica_Meta_3_ciclo_1.pdf

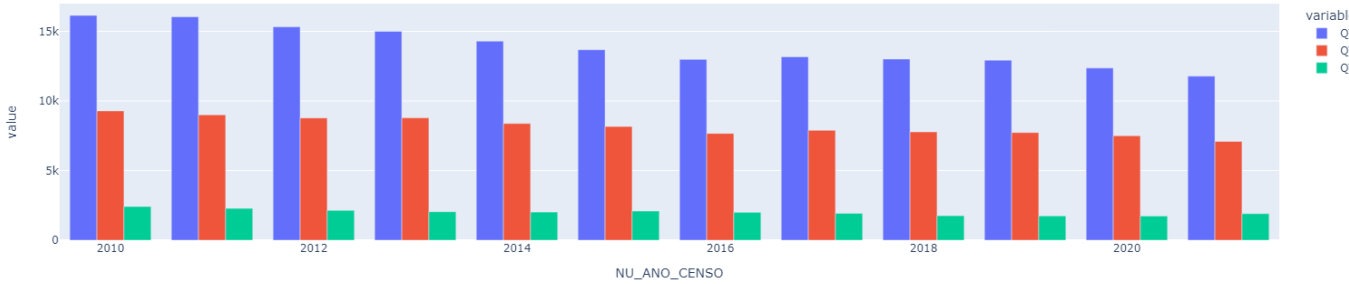

Figura 4 – Quantidade de Matrículas por ano e grau do Município de Moreno-PE

5 Considerações Finais e Trabalhos Futuros

O presente artigo teve o objetivo de apresentar esse conjunto de dados que processado sobre os dados disponibilizados pelo INEP em sua nova versão de 2022, onde é possível termos os dados históricos de 2010 até 2021,o último ano disponível até a escrita deste material. Os dados foram reorganizados de forma a que fosse possível a usabilidade de uma quantidade menor de arquivos a depender do objetivo do utilizador. Também é fornecido um dicionário atualizado que reflete a estrutura atual do conjunto de dados.

Algumas limitações podem ser observadas sobre dados ausentes no conjunto de dados, porém nessa etapa do projeto o objetivo é fornecer os dados mais próximos aos disponibilizados pelo INEP e posteriormente ser ajustado a depender da necessidade de cada projeto.

Sobre trabalhos futuros podemos destacar a próxima etapa que seja o desenvolvimento de uma *API (Application Programming Interface)* para o fornecimento desses dados utilizando um banco de dados relacional. Além disso, busca-se utilizar esses dados em painéis de acompanhamento das metas relativas ao PNE.

Referências

BALBINOT, A. D.; HAUBERT, A. Análise temporal das matrículas em educação especial entre 2005 e 2013 no estado do paraná. *Revista Prâksis*, Universidade Feevale, v. 2, p. 121–132, 2015. Citado na página 18.

BALBINOT, A. D.; HAUBERT, A. Análise de matrículas como indicadores da evolução da educação especial no estado do rio de janeiro. *REVISTA ELETRÔNICA PESQUISEDUCA*, v. 9, n. 19, p. 663–673, 2017. Citado na página 18.

BARRETO, I. M. de S.; FREITAS, A. E. S. Gerando inteligência através de microdados: uma proposta de business intelligence para a área de ensino do instituto federal da bahia (ifba). 2020. Citado na página 9.

CONTE, V. d. S. Mineração de dados educacionais para avaliar os fatores que influenciam no desempenho de candidatos do enem. 2019. Citado na página 11.

DINIZ, E. O censo escolar. *Revista Brasileira de Estudos Pedagógicos*, v. 80, n. 194, 1999. Citado na página 9.

FERREIRA, J. et al. O processo etl em sistemas data warehouse. In: *INForum*. [S.l.: s.n.], 2010. p. 757–765. Citado na página 13.

FILHO, J. A. C.; ISOTANI, S.; PENTEADO, B. E. Utilização de notas escolares para predição da nota enem em ciências humanas. 2021. Citado na página 11.

FRENEDA, F. C. B. et al. Múltiplos fatores do desempenho escolar: uma análise dos microdados do inep sobre a educação no distrito federal. Universidade Católica de Brasília, 2020. Citado na página 9.

GONÇALVES, M. V. F. et al. Datasets curados e enriquecidos com proveniência da campanha nacional de vacinação contra covid-19. In: SBC. *Anais do III Dataset Showcase Workshop*. [S.l.], 2021. p. 148–159. Citado na página 11.

JAMES, E. A.; MILENKIEWICZ, M. T.; BUCKNAM, A. *Participatory action research for educational leadership: Using data-driven decision making to improve schools*. [S.l.]: Sage, 2008. Citado na página 9.

LAI, M. K.; SCHILDKAMP, K. Data-based decision making: An overview. *Data-based decision making in education*, Springer, p. 9–21, 2013. Citado na página 9.

PINTO, G. da S.; JÚNIOR, O. d. G. F.; COSTA, E. de B. Identificação dos fatores de melhorias no ideb pelo uso de mineração de dados: Um estudo de caso em escolas municipais de teotônio vilela-alagoas. *RENOTE*, v. 17, n. 3, p. 183–193, 2019. Citado na página 11.

VASCONCELOS, F. F. et al. Candidata: um dataset para análise das eleições no brasil. In: SBC. *Anais do III Dataset Showcase Workshop*. [S.l.], 2021. p. 160–168. Citado na página 11.

WANDERLEY, P. F. et al. Uso de business intelligence para avaliação de indicadores de desempenho na educação básica: um estudo de caso no estado do acre. Universidade Federal de Campina Grande, 2021. Citado na página 11.