

Banco de Dados Educacionais da Cátedra Sérgio Henrique Ferreira

Carlos Eduardo Rodrigues Dos Santos¹

Cátedra Sérgio Henrique Ferreira do Instituto de Estudos Avançados da USP de Ribeirão Preto
Ribeirão Preto, São Paulo, Brasil
carloredrs@usp.br

Pedro Sartori Dias Dos Reis²

Cátedra Sérgio Henrique Ferreira do Instituto de Estudos Avançados da USP de Ribeirão Preto
Ribeirão Preto, São Paulo, Brasil
pedro.sartori@usp.br

Rafael Naime Ruggiero³

Departamento de Psicobiologia da Universidade Federal de São Paulo (UNIFESP) e Cátedra Sérgio Henrique Ferreira do Instituto de Estudos Avançados da USP de Ribeirão Preto
São Paulo, São Paulo, Brasil
rafaruggiero@usp.br

Mozart Neves Ramos⁴

Cátedra Sérgio Henrique Ferreira do Instituto de Estudos Avançados da USP de Ribeirão Preto
Ribeirão Preto, São Paulo, Brasil
mozartnramos@gmail.com

Resumo

O Banco de Dados Educacionais criado na Cátedra Sérgio Henrique Ferreira tem como objetivo centralizar informações sobre a educação pública brasileira, abrangendo diversas modalidades e etapas de ensino ao longo dos anos. Ele reúne dados desde coordenadas e informações específicas de cada unidade escolar até indicadores educacionais gerados pelo INEP (*Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira*). Essas informações estão organizadas por instâncias – Escolas, Municípios, Estados, Regiões e Brasil – e métricas (indicadores) sobre essas instâncias – permitindo análises refinadas e comparáveis entre diferentes níveis de agregação. O banco de dados se destina a apoiar políticas educacionais e estudos acadêmicos, fornecendo uma base para investigações sobre o desempenho educacional, condições estruturais das escolas e, principalmente, suas desigualdades dentro da rede.

Palavras-chave: Banco; Dados; Educação; Análise; Políticas.

¹ Os autores agradecem ao Santander Universidades, à B3 Social e à Fundação Telefônica pelo apoio financeiro ao trabalho da Cátedra Sérgio Henrique Ferreira do IEA da USP de Ribeirão Preto.

²

³

⁴

1 Introdução

A Cátedra Sérgio Henrique Ferreira, vinculada ao Instituto de Estudos Avançados da Universidade de São Paulo – Polo Ribeirão Preto, tem como objetivo o estudo das desigualdades educacionais e a promoção de políticas públicas que contribuam para sua redução. Com o aumento da demanda por projetos de pesquisa por parte de diversos parceiros institucionais, identificou-se a necessidade de maior agilidade nos processos de coleta, organização e análise de dados educacionais.

Diante desse cenário, foi concebido o projeto do Banco de Dados Educacionais, com a finalidade de centralizar e estruturar informações sobre escolas brasileiras em um formato acessível e comparável. Para tanto, desenvolveu-se uma modelagem relacional utilizando o Sistema de Gerenciamento de Banco de Dados (SGBD) *PostgreSQL* (POSTGRES, 2025), permitindo a integração de dados provenientes de diferentes fontes. A base de dados é composta, principalmente, pelos microdados do Censo Escolar e da Prova Brasil, disponibilizados pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP).

Considerando que, ao longo dos anos, esses dados frequentemente não seguem um padrão uniforme, foi necessário realizar etapas de limpeza, padronização e complementação das informações. Para as escolas cujas coordenadas geográficas estavam ausentes, recorreu-se à API do *Google Maps* (GOOGLE, 2025), a fim de garantir a completude espacial dos dados. Todo o conteúdo foi modelado de modo a permitir análises em diferentes níveis de agregação — escolas, municípios, estados e regiões —, bem como por etapa de ensino e localização geográfica.

2 Metodologia

A Metodologia deste trabalho compreende a modelagem, a coleta de dados das instâncias, indicadores e coordenadas, e a respectiva população do banco.

2.1 – Modelagem:

Para modelar o banco de dados, inicialmente foi utilizada a ferramenta de modelagem do programa *MySQL Workbench*. Nesse ponto, a modelagem foi estruturada para tratar todas as instâncias da mesma forma, de modo que fosse possível compará-las entre si e recuperar informações de modo eficiente. É necessário ressaltar que no banco de dados *PostgreSQL* – ferramenta onde o banco de dados foi implementado –, atributos categóricos das instâncias (ex: Tipo de etapa escolar, dependência administrativa, etc.) foram tratados com o tipo “ENUM” (*Enumerated Types*)(POSTGRES), que por sua vez não existe dentro da ferramenta *MySQL*, e portanto, foi representada como o tipo “TEXT”.

2.2 – Coleta de Dados

2.2.1 – Coleta dos dados das escolas:

Para coletar os dados das escolas, foram utilizadas as planilhas dos Microdados do Censo Escolar, disponibilizadas pelo INEP, abrangendo o período de 2005 a 2023 — intervalo em que muitos dos indicadores educacionais atualmente utilizados passaram a ser coletados de forma sistemática. Como algumas escolas podem ter aparecido apenas em determinados anos dentro desse período, foi necessário adotar uma estratégia que garantisse sua inclusão no banco de dados, mesmo que não tivessem informações em todos os anos. Para isso, foi realizada uma operação de cruzamento entre os dados de diferentes anos, de forma que todas as escolas que constaram ao menos uma vez no Censo fossem consideradas. Essa abordagem assegurou a abrangência do banco e evitou a perda de instituições relevantes para as análises posteriores.

2.2.2 – Coleta de dados dos estados e regiões:

Para coletar dados dos municípios, estados, regiões e do Brasil, se fez necessário a utilização de dados do IBGE (Instituto Brasileiro de Geografia e Estatística), como o código padrão de cada Unidade Federativa e também de seus respectivos municípios. O código do Brasil foi usado de acordo com a Organização Internacional para Padronização (ISO), através da norma ISO 3166-1, sendo que o código referente ao Brasil é o 076.

2.2.3 – Coleta de dados de coordenadas faltantes no INEP:

Para algumas escolas, dados sobre localização estavam ausentes nas planilhas do INEP. Para contornar esta limitação, foi necessária a utilização de uma ferramenta específica para obter coordenadas geográficas a partir de um endereço. Em nosso caso, a ferramenta utilizada foi a *Geocoding API*, do *Google*. A partir disso, foi gerada uma lista com os endereços das escolas cujas coordenadas eram desconhecidas, e logo em seguida foram feitas buscas das coordenadas, com alto nível de precisão. Para fazer a validação de que as coordenadas dessas escolas correspondessem com a realidade, foi feito um teste plotando os pontos sobre o mapa do Brasil, onde foi possível notar que a localização dos pontos gerados por coordenadas se encontrava dentro das dependências dos estados e municípios correspondentes.

2.3 – População do Banco:

Para popular o banco, foi necessário primeiramente agrupar todos os dados de indicadores, de modo que para cada indicador houvesse apenas um arquivo em formato CSV (*Comma Separated Values*) contendo todos os anos em que aquele indicador foi coletado. Após fazer essa união, foi necessário formatar cada planilha para que o resultado fosse compatível com o formato projetado em cada tabela da modelagem. Todos os dados foram tratados usando as linguagens de programação *Python* e *R*, além de *SQL* para tratar do banco. Foi adotado o uso de *Containers* através do *Docker Engine*, ferramenta que possibilitou criar e hospedar o banco de dados em um ambiente controlado para testes e desenvolvimento, uma vez que permite que os desenvolvedores trabalhem em conjunto remotamente e independente da configuração de suas máquinas de trabalho divergirem. Aqui, é válida a menção ao software *Dbeaver*, que foi utilizado como meio para se conectar ao SGBD escolhido (PostgreSQL).

3 Resultados

A implementação do Banco de Dados Educacionais na Cátedra Sérgio Henrique Ferreira apresenta resultados nas pesquisas educacionais da Cátedra Sérgio Henrique Ferreira. A centralização dos dados e a modelagem relacional permitiram a execução de estudos complexos de maneira mais rápida e eficiente. Por exemplo, pesquisas voltadas ao desempenho educacional em diferentes regiões do Brasil foram realizadas de forma otimizada, graças à disponibilidade imediata de dados escolares consolidados.

Ao concentrar os indicadores educacionais, como taxas de aprovação, abandono e desempenho em avaliações, o banco permitiu que realizássemos análises multivariadas robustas. Isso resultou em *insights* valiosos sobre os fatores que mais influenciam o desempenho escolar em diferentes contextos, o que seria difícil de alcançar sem um sistema de dados centralizado.

Na atual etapa, o banco de dados encontra-se populado com dados públicos até o ano de 2023, porém ainda não foi disponibilizado para uso aberto ao público, sendo usado apenas para uso interno na Cátedra, sendo necessária uma plataforma para acesso externo. Através deste banco, diversos projetos e estudos vem sendo desenvolvidos, como por exemplo o *Dashboard* Interativo feito para o município de São Paulo, e que, utilizando dados extraídos deste banco de dados, cria visualizações em que o usuário pode personalizar de acordo com seus interesses.

3.1 Discussão

A implementação do Banco de Dados Educacionais na Cátedra Sérgio Henrique Ferreira representou um avanço estratégico diante do aumento progressivo na demanda por análises e evidências para embasar políticas públicas educacionais. Diante da crescente complexidade dos projetos e do volume de dados utilizados, tornou-se essencial dispor de uma estrutura centralizada, padronizada e eficiente para armazenamento e análise de informações educacionais.

A centralização dos dados escolares e a adoção de uma modelagem relacional robusta transformaram significativamente a capacidade de execução de estudos complexos. Antes limitadas por processos manuais e fragmentação das fontes de dados, as análises agora podem ser realizadas de forma ágil e integrada, o que ampliou o escopo e a sofisticação das investigações conduzidas pela Cátedra.

Dois exemplos ilustram bem o impacto dessa transformação. O primeiro é o projeto desenvolvido em parceria com a Secretaria de Educação de Jundiaí, no qual o banco de dados foi utilizado para analisar o desempenho escolar de unidades que implementaram uma nova metodologia de ensino. Graças à consolidação e organização prévia dos indicadores educacionais para o ano de 2023, foi possível conduzir rapidamente uma avaliação robusta dos efeitos da política adotada, economizando tempo e recursos.

Outro exemplo é o estudo conduzido no estado de Alagoas, onde o banco viabilizou a realização de análises exploratórias de desempenho escolar por meio de visualizações como mapas de calor e gráficos interativos. Essa etapa, que

anteriormente exigiria esforço considerável de coleta e organização, pôde ser realizada com muito mais agilidade e precisão, contribuindo para uma compreensão mais clara das desigualdades regionais no estado.

Apesar dos avanços, existem desafios a serem superados. Atualmente, o banco está populado com dados públicos até o ano de 2023, sendo atualizado progressivamente à medida que novos dados são divulgados. Seu uso, por ora, está restrito ao ambiente interno da Cátedra, uma vez que o projeto ainda se encontra em desenvolvimento e há preocupações com a segurança da infraestrutura. Por esse motivo, está em desenvolvimento uma plataforma que permitirá o acesso externo ao banco de dados, com os devidos cuidados quanto à segurança, à privacidade dos dados e à usabilidade por parte de diferentes perfis de usuários.

A disponibilização pública dessa interface será um passo crucial para o amadurecimento do projeto, potencializando seu impacto ao permitir que outras instituições e pesquisadores possam também se beneficiar da estrutura criada. A continuidade desse processo depende, portanto, não apenas do aperfeiçoamento técnico, mas também da definição de protocolos claros de governança e acesso.

4 Conclusões e Considerações Finais

A criação do Banco de Dados Educacionais pela Cátedra Sérgio Henrique Ferreira representa uma resposta concreta às crescentes demandas por agilidade, integração e profundidade nas análises de dados educacionais. Ao consolidar informações públicas provenientes do Censo Escolar e de outras bases do INEP em uma estrutura relacional robusta, o projeto viabilizou a realização de estudos mais rápidos, precisos e abrangentes. A incorporação de dados geográficos, aliada à padronização temporal, ampliou as possibilidades analíticas e tem contribuído para subsidiar decisões baseadas em evidências em diferentes contextos.

Ainda que o banco esteja, por ora, restrito ao uso interno da Cátedra, os resultados obtidos demonstram seu potencial como infraestrutura de apoio à pesquisa e à formulação de políticas públicas educacionais. A futura disponibilização pública da plataforma representa uma etapa estratégica para expandir o impacto do projeto, promovendo o uso qualificado de dados por pesquisadores, gestores e demais interessados no enfrentamento das desigualdades educacionais no Brasil.

5 Referências

MCKINNEY, et-al (2010). Data structures for statistical computing in python. In Proceedings of the 9th Python in Science Conference (Vol. 445, pp. 51–56).

JUDITH BOWMAN, et-al (2001). The Practical SQL Handbook Using SQL Variants. Fourth Edition. ISBN 0-201-70309-2.

KNUTH, J. C. *The Programming of Computer Art*. Stanford, California, 1988. This is a full BOOKLET entry.

DOCKER, Inc. Docker Documentation. Disponível em: <https://docs.docker.com/>. Acesso em: 28 abr. 2025.

POSTGRES. Disponível em: <https://www.postgresql.org/>. Acesso em: 28 abr. 2025.

G
O
O
G
L
E
.

G
e
o
c
o
d
i
n
g

A
P
I

o
v
e
r
v
i
e
w
.

D
i
s
p
o
n
í
v
e
l

e
m
:

h
t
t
p
s
:
/