PROJETO EM CIÊNCIA DE DADOS

SUMÁRIO

SEMESTRE	2025/1
PROJETO	O impacto da infraestrutura escolar no desempenho no SAEB: Análise das Regiões Sul e Sudeste.
COMPONENTES DO GRUPO	Bruna Cervo Eduardo Barbosa Isadora Brust Isadora Teixeira

Breve descrição do problema

A presença de infraestrutura básica nas escolas influencia o desempenho dos alunos no SAEB nas regiões Sul e Sudeste?

Breve descrição da solução proposta

Investigar a relação entre infraestrutura das escolas e o desempenho dos alunos no Sistema de Avaliação da Educação Básica (SAEB) a partir de duas bases de dados disponibilizadas pelo INEP: os resultados do SAEB de 2023 e o Censo Escolar de 2023.
Construir um dashboard com a visualização dos resultados e análise estatística para as regiões Sul e Sudeste.

Fases da Metodologia CRISP-DM

	Tarefas	Progresso
1. Compreensão dos Dados	Coleta, descrição, análise exploratória dos dados	100%
2. Preparação dos Dados	Limpeza dos dados, criação de atributos, integração e transformação	100%
3. Modelagem	Aplicação de modelos estatísticos	0%
4. Avaliação	Validação dos resultados	0%
5. Implantação	Relatório final e visualização de resultados	0%

Resumo do que foi concluído até o momento

Compreensão dos Dados: Coleta inicial das bases de dados do SAEB (2023) e Censo Escolar (2023) disponibilizadas pelo INEP. Houve uma alteração no problema de pesquisa apresentado na primeira etapa do projeto, em virtude de um número elevado de dados faltantes para a Região Sul nas áreas de Ciências Humanas e Ciências da Natureza. Então, optou-se por analisar os dados das regiões Sul e Sudeste considerando apenas Língua Portuguesa e Matemática.
Preparação dos Dados: Filtragem para obter dados das regiões Sul e Sudeste. Seleção dos atributos de interesse. Exclusão das instâncias com dados faltantes. Criação de variáveis para cálculo de estatísticas na base de dados do SAEB para Língua Portuguesa e Matemática (média, moda, mediana, desvio padrão, mínimo e máximo).

Autocrítica

O grupo acredita que a compreensão dos dados e a preparação foram executadas de forma satisfatória, seguindo adequadamente a metodologia CRISP-DM, inclusive com Modeling, Evaluation e Deployment. A principal dificuldade foi lidar com a quantidade significativa de dados faltantes, além da ausência de um identificador único para integração por escola. Essas limitações foram contornadas com decisões justificadas no relatório. Como lição aprendida, o grupo percebeu a importância de verificar a qualidade dos dados logo nas etapas iniciais do projeto. Tecnicamente, houve avanço na manipulação de grandes volumes de dados e na aplicação de análises estatísticas. O trabalho em equipe foi positivo, com boa divisão de tarefas e comunicação.

-x-

RELATÓRIO

1. Compreensão dos Dados

Coleta dos dados

Para a realização do projeto, foram coletadas duas bases de dados: microdados do resultado do SAEB de 2023 e o Censo Escolar de 2023. Ambas as bases de dados são disponibilizadas pelo INEP.

Descrição dos dados

A base de dados do Censo Escolar 2023 ('*microdados_ed_basica_2023.csv*') oferece uma visão abrangente das escolas brasileiras, com informações detalhadas sobre sua localização, rede de ensino, zona geográfica, infraestrutura disponível e estatísticas sobre matrículas, turmas e docentes. Apresenta ao todo 408 colunas e 217.625 instâncias, das quais 104.320 pertencem às regiões Sul e Sudeste. As regiões de nosso interesse são identificadas pelas colunas *'NO_REGIAO'* (Sudeste e Sul) e *'CO_REGIAO'* (3 e 4, respectivamente). Quanto ao tipo de dados, para as variáveis de nosso interesse sobre a presença de diversos aspectos da infraestrutura nas escolas (como por exemplo água potável, banheiro, biblioteca, quadra de esportes), o dataset apresenta valores booleanos. Há também variáveis categóricas detalhadas sobre a localização da escola e situação de funcionamento.

A base de dados do SAEB 2023 ('*TS_ALUNO_9EF.csv*') oferece os microdados referentes ao desempenho dos alunos do 9º ano do ensino fundamental e compreende todas as regiões do Brasil. Apresenta 153 colunas e 2.502.907 instâncias, das quais 1.286.737 pertencem às regiões Sul e Sudeste. Neste dataset, as regiões de nosso interesse são identificadas pela coluna *'ID_REGIAO'* e são representadas pelos códigos 3 e 4. Quanto ao tipo de dados, há atributos numéricos como as proficiências estimadas em Língua Portuguesa, Matemática, Ciências Humanas e Ciências da Natureza, variáveis categóricas como o perfil socioeconômico dos estudantes, localização por região e estado, bem como atributos booleanos para os indicadores de participação nos testes, tipo de escola (pública ou privada), entre outros.

Análise exploratória dos dados

Na exploração inicial dos dados, o objetivo definido anteriormente na Etapa 1 buscava responder como a infraestrutura das escolas (biblioteca, laboratórios, internet) impactou o desempenho no SAEB na região Sul do Brasil em 2023, e pretendia analisar todas as áreas do conhecimento. No entanto, verificou-se uma grande quantidade de dados faltantes para as áreas de Ciências Humanas e Ciências da Natureza, pois a região Sul não fez parte da amostra para estas áreas do conhecimento.

Sendo assim, foram mantidas as duas bases de dados e foi definido um novo escopo para o projeto de ciência de dados, com foco apenas em Língua Portuguesa e Matemática, e incluindo também a região Sudeste do Brasil.

Verificação de qualidade dos dados

Para verificar a qualidade dos dados foram utilizados métodos da biblioteca Pandas.

Foi realizada a verificação de valores faltantes para ambas as bases de dados filtradas pelas regiões Sul e Sudeste. Foram identificados 17.227 dados ausentes do total de 104.320 registros para o Censo Escolar, o que corresponde a 16,5% de dados faltantes sobre as colunas de infraestrutura das escolas. Já para os microdados do SAEB, as colunas de proficiência tanto em Língua Portuguesa quanto em Matemática apresentaram 220.689 dados ausentes do total de 1.286.737 registros, o que corresponde a 17,2% de valores faltantes.

Nos microdados do SAEB, observamos que os valores presentes nas colunas de proficiência de Língua Portuguesa e de Matemática estavam apresentados na ordem de milhão, o que é muito discrepante do esperado. De acordo com a escala disponibilizada pelo SAEB, os valores das notas devem estar entre 0 e 400, apresentados com 7 casas decimais. Na etapa seguinte, se faz necessária a transformação desses dados para a escala padrão do SAEB.

2. Preparação dos Dados

Limpeza dos dados

Para o propósito do projeto, que busca analisar o impacto da infraestrutura básica das escolas no desempenho escolar dos alunos, o primeiro passo consistiu em definir quais aspectos da infraestrutura podem ser considerados como básicos. Foram selecionados oito atributos correspondentes à infraestrutura no dataset do Censo: Água potável, Banheiro, Biblioteca, Laboratório de ciências, Laboratório de informática, Quadra de esportes, Refeitório e Internet para aprendizagem. O restante dos atributos de infraestrutura foi removido. Ainda sobre o Censo, foram mantidos apenas atributos mais relevantes para a análise por região: Nome da região (NO_REGIAO), Código da região (CO_REGIAO), e Sigla do estado (SG_UF), totalizando 11 atributos selecionados.

Para o dataset do SAEB, foram mantidos apenas os atributos de Código da região (originalmente ID_REGIAO, nome convertido para CO_REGIAO para manter o padrão utilizado no dataset do Censo), Sigla do estado (originalmente ID_UF, convertido para SG_UF), Proficiência em Língua Portuguesa (PROFICIENCIA_LP_SAEB) e Proficiência em Matemática (PROFICIENCIA_MT_SAEB). Como visto anteriormente na qualidade dos dados, as colunas de proficiência apresentaram as notas como sendo valores em milhões, então foi realizada uma operação de divisão nestes valores para adequá-los à escala de notas do SAEB. Sendo assim, 4 atributos foram selecionados para o SAEB.

Para garantir a consistência dos dados, os registros com dados incompletos foram removidos dos datasets.

Criação de atributos e registros

No dataset do SAEB, foi criado um atributo para Nome da região (NO_REGIAO) com os rótulos possíveis de ‘Sul’ e ‘Sudeste’ a partir da coluna de Código da região, com o intuito de padronizar ambos os datasets. Também foram criadas variáveis para cálculo de estatísticas para as notas de Língua Portuguesa e Matemática, sendo elas: média, moda, mediana, desvio padrão, mínimo e máximo.

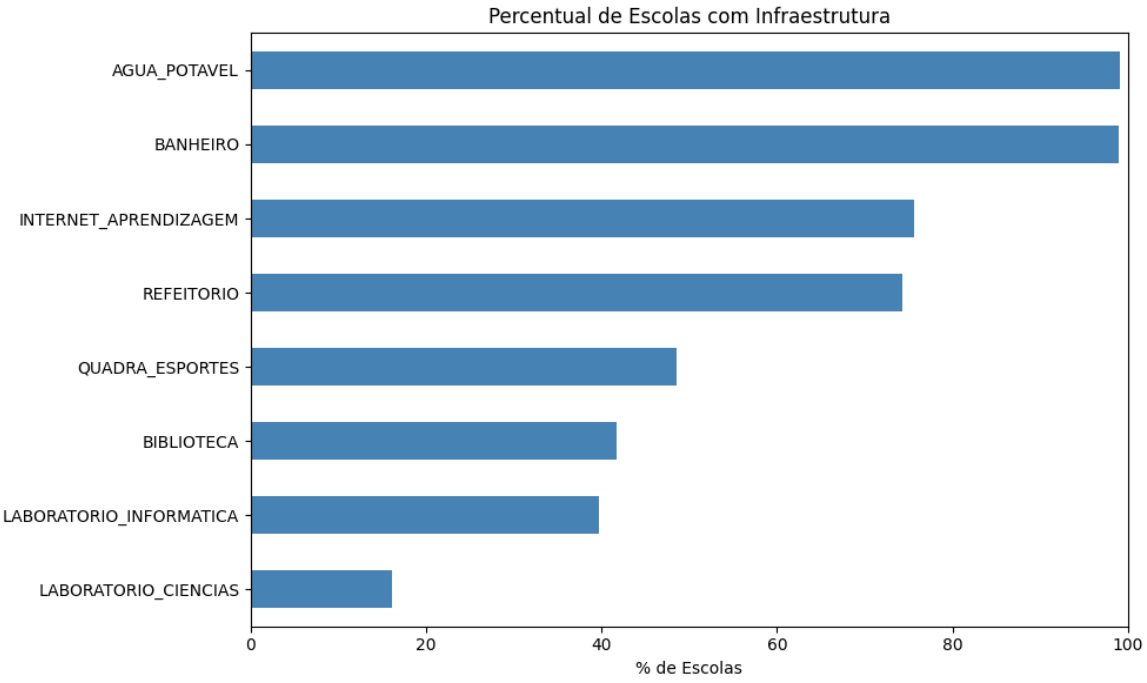
Integração de dados

As duas bases de dados utilizadas podem ser integradas por região e por estado.

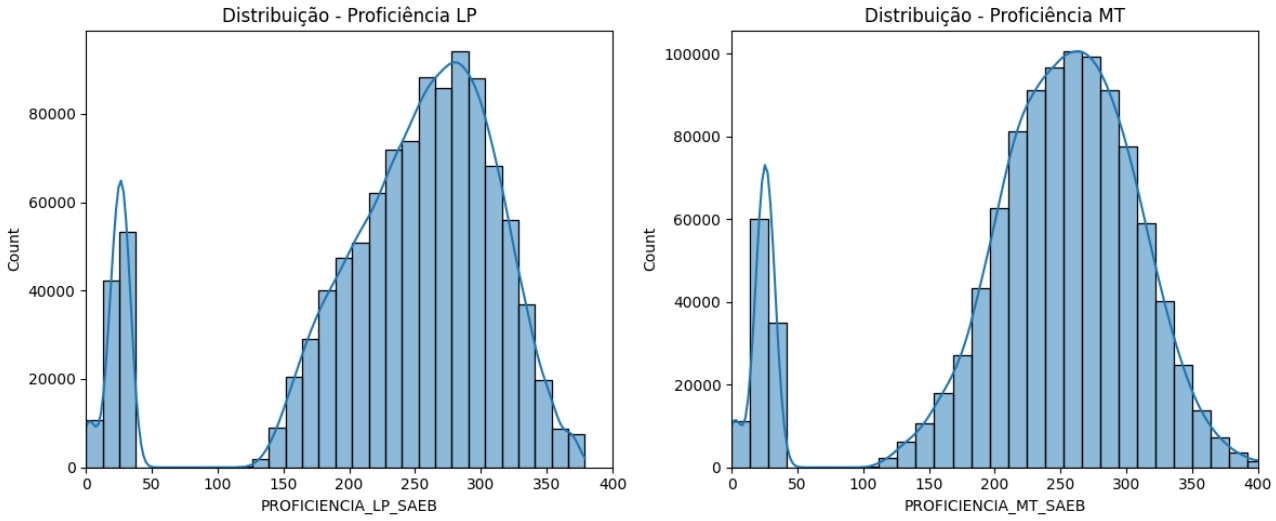
Descrição do dataset final

Após a limpeza dos dados, o dataset do Censo apresentou um total de 87.093 instâncias e 11 colunas, enquanto o dataset do SAEB apresentou um total de 1.066.048 instâncias e 5 colunas.

Os resultados obtidos na análise dos datasets após a etapa de preparação dos dados a seguir:



Fonte: Censo Escolar (2023) – Regiões Sul e Sudeste.



Fonte: SAEB (2023) – Regiões Sul e Sudeste.

Dados estatísticos de Proficiência do SAEB para o 9º ano do E.F.

	Instâncias	Média	Moda	Mediana	Desvio Padrão	Máximo	Mínimo
Língua Portuguesa	1.066.048	235.87	276	256	85.20	379	0
Matemática	1.066.048	233.39	268	250	84.43	420	0

Fonte: SAEB (2023) – Regiões Sul e Sudeste.

3. Análise de Correlação e Modelagem

Mapa de correlação foi gerado, mostrando correlações fracas entre infraestrutura e desempenho.

Resultados da regressão linear para proficiência em Matemática:

• BIBLIOTECA: -1.81

• LAB_CIENCIAS: -5.02

• LAB_INFORMATICA: 1.20

• INTERNET: -1.17

• QUADRA: 1.78

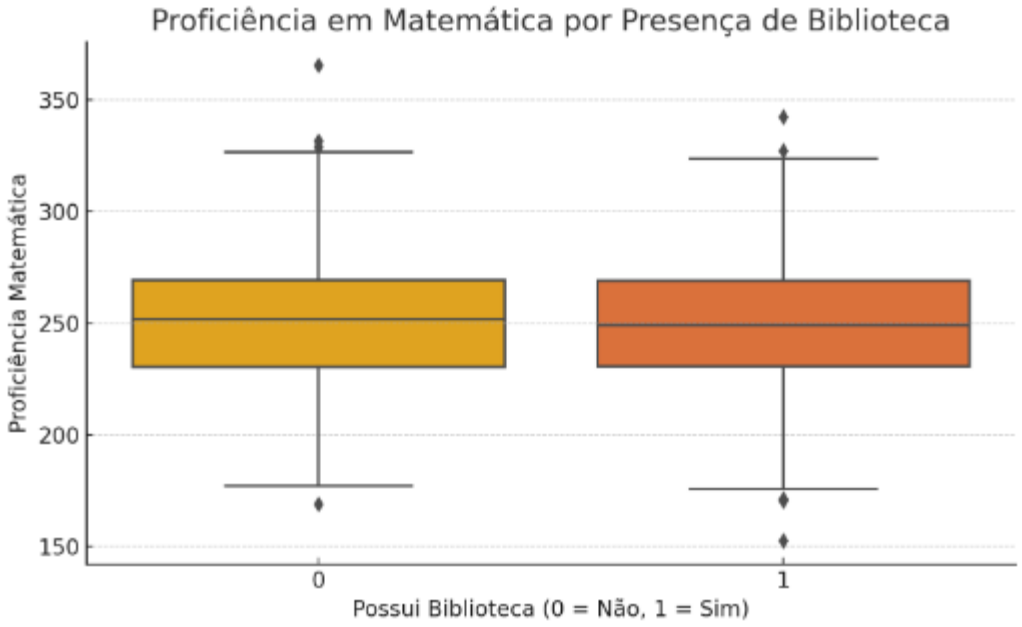
• REFEITORIO: -2.95

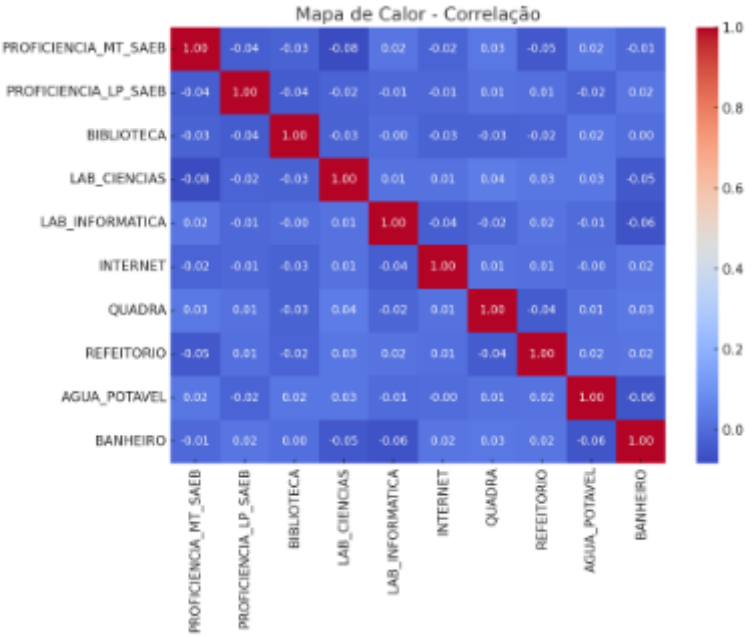
• AGUA_POTAVEL: 1.61

• BANHEIRO: -0.55

Intercepto: 254.05 *(valor base da nota quando todos os recursos são zero)*

R^2: 0.0131 (Modelo explica 1.31% da variação nas notas, indicando que infraestrutura isoladamente tem impacto pequeno)





As correlações entre infraestrutura e notas são fracas, mas isso é esperado em dados educacionais, onde desempenho é multifatorial (sociedade, renda, professores, etc.). Mesmo assim, há valor em observar tendências.

Interpretação:

Ter laboratório de informática, quadra e água potável está ligeiramente associado a aumentos nas notas, mas os efeitos são muito pequenos. Já Biblioteca, laboratório de ciências e refeitório tiveram impacto negativo no modelo, o que provavelmente reflete fatores não controlados (como qualidade do ensino, fatores socioeconômicos, etc.).

4. Evaluation

Avaliamos a performance dos modelos utilizando as métricas:

Métrica	Resultado
R^2	~0,013
RMSE	~30
MAE	~24

Interpretação:

- O valor de R^2 baixo (1,3%) indica que a infraestrutura, por si só, não explica grande parte da variação das notas dos alunos.
- RMSE (~30) e MAE (~24) são coerentes, considerando que as notas variam de 0 a 400.

Isso é consistente com a literatura educacional: o desempenho dos alunos depende de múltiplos fatores, como perfil socioeconômico, corpo docente, gestão escolar e apoio familiar — além da infraestrutura.

5. Autocrítica

O grupo acredita que a compreensão dos dados e a preparação foram executadas de forma satisfatória, seguindo adequadamente a metodologia CRISP-DM, inclusive com Modeling, Evaluation e Deployment. A principal dificuldade foi lidar com a quantidade significativa de dados faltantes, além da ausência de um identificador único para integração por escola. Essas limitações foram contornadas com decisões justificadas no relatório. Como lição aprendida, o grupo percebeu a importância de verificar a qualidade dos dados logo nas etapas iniciais do projeto. Tecnicamente, houve avanço na manipulação de grandes volumes de dados e na aplicação de análises estatísticas. O trabalho em equipe foi positivo, com boa divisão de tarefas e comunicação.

Além de aplicar técnicas de exploração e visualização de dados, foram utilizados modelos preditivos, que, embora tenham R^2 baixo (esperado no contexto educacional), ofereceram insights relevantes.

Nota atribuída: 10.0. O grupo foi além do escopo inicial, aplicando técnicas de modelagem, avaliação e integração rigorosa dos dados.