

## PROJETO EM CIÊNCIA DE DADOS

### SUMÁRIO

SEMESTRE	2025/1
PROJETO	Trabalho Final
COMPONENTES DO GRUPO	Eduardo Duraes Eduardo Santos Leonardo Avila Lampert

#### Breve descrição do problema

O grupo investigará a possível relação entre a infraestrutura das escolas brasileiras e o desempenho médio dos alunos no Exame Nacional do Ensino Médio (ENEM), utilizando dados referentes ao ano de 2015. A hipótese central é de que escolas com melhor infraestrutura tendem a apresentar médias mais altas, o que pode revelar desigualdades estruturais no sistema educacional.

#### Breve descrição da solução proposta

O grupo calculou a média geral do ENEM por escola com base nas cinco áreas avaliadas. Em seguida, criou quatro medidas de infraestrutura: presença de itens (INFRA\_BOOL\_SOMA), quantidade de equipamentos (QT\_EQUIP\_TOTAL), número de funcionários (QT\_FUNC\_TOTAL) e itens críticos (INFRA\_CRITICA\_SOMA), além da formação docente. Por fim, analisou-se a correlação entre essas variáveis e a média do ENEM para investigar possíveis relações. As entregas incluem os cálculos, as medidas definidas e a análise de correlação.

#### Fases da Metodologia CRISP-DM

Fase	Tarefas realizadas	Percentual de conclusão
1. Entendimento dos dados	Leitura e análise dos dados do ENEM e da infraestrutura escolar; identificação de variáveis relevantes.	100%
2. Preparação dos dados	Limpeza, filtragem, criação de médias e construção das variáveis INFRA_BOOL_SOMA, INFRA_CRITICA_SOMA, etc.	100%
3. Modelagem	Aplicação de análise de correlação para investigar relações entre variáveis.	100%
4. Avaliação	Interpretação dos resultados das correlações; análise da relevância dos indicadores criados.	100%

## Resumo do que foi concluído até o momento

O grupo finalizou a preparação dos dados, o cálculo das médias do ENEM por escola e a análise de correlação com variáveis de infraestrutura e formação docente. Os resultados mostram que, isoladamente, a infraestrutura tem pouca influência no desempenho. A formação dos professores apresentou a maior correlação ( $\sim 0,264$ ), mas ainda assim considerada fraca. Isso indica que fatores individuais dos alunos podem ter papel mais relevante. As atividades seguiram conforme o planejado, restando apenas ajustes finais.

## Autocrítica

O grupo acredita que, até o momento, o trabalho está sendo conduzido de forma satisfatória, com boa aderência à metodologia CRISP-DM. Todas as etapas até a avaliação foram seguidas conforme o proposto, com organização e divisão de tarefas bem definidas. Do ponto de vista técnico, aprendemos a lidar com grandes volumes de dados educacionais, a criar variáveis relevantes e aplicar métodos estatísticos como a correlação. Do ponto de vista sociocomportamental, o trabalho em grupo exigiu organização, comunicação constante e colaboração para superar dúvidas e dificuldades técnicas.

Damos ao grupo a nota 9,0, pois cumprimos grande parte do escopo com qualidade, mas reconhecemos que ainda há espaço para melhorar a interpretação dos resultados e refinar a apresentação final.

# RELATÓRIO

## 1. Compreensão dos Dados

### Coleta dos dados

Os dados foram obtidos dos microdados públicos do INEP (ENEM por escola e Censo Escolar de 2015). A coleta foi feita por download direto do site oficial. O maior desafio foi a dimensão das bases e a necessidade de selecionar somente colunas relacionadas à infraestrutura e desempenho.

### Descrição dos dados

Dados booleanos (ex: tem ou não biblioteca, quadra, etc.), que foram somados para gerar métricas como `INFRA_BOOL_SOMA`. Dados numéricos de infraestrutura (quantidade de equipamentos, funcionários), notas por área do ENEM, com uma média geral criada pelo grupo, códigos de identificação para cruzamento das bases.

### Análise exploratória dos dados

Foram eliminadas colunas não relacionadas à infraestrutura. Criamos métricas somando os dados booleanos, os equipamentos e os funcionários por escola. Calculamos a média geral do ENEM e aplicamos correlação entre essa média e as variáveis.

### Verificação de qualidade dos dados

Os dados estavam bem estruturados. Foi encontrada apenas uma pequena quantidade de valores nulos (19) na coluna de formação docente, sem impacto significativo na análise.

## 2. Preparação dos Dados

### Limpeza dos dados

A limpeza foi feita com base no dicionário de dados, removendo todas as colunas que não estavam relacionadas à infraestrutura escolar. Também foram excluídas 19 linhas com valores nulos na coluna FORMACAO\_DOCENTE, já que representavam uma pequena parte do total.

### Criação de atributos e registros

Foram criadas colunas para facilitar a análise. A MEDIA\_GERAL foi calculada a partir da média das cinco notas do ENEM por escola. Também criamos indicadores de infraestrutura: INFRA\_BOOL\_SOMA (soma dos recursos booleanos), INFRA\_CRITICA\_SOMA (saneamento básico), QT\_EQUIP\_TOTAL (equipamentos) e QT\_FUNC\_TOTAL (docentes + funcionários). Esses atributos ajudam a quantificar as condições de cada escola.

### Integração de dados

Os dados do ENEM por escola foram integrados com os dados do Censo Escolar 2015 usando o código identificador da escola. Apenas as colunas relevantes para a análise foram mantidas. Durante a integração, colunas redundantes ou sem relação com infraestrutura foram eliminadas.

### Descrição do dataset final

O dataset final utilizado na etapa de modelagem é o resultado da integração entre os dados do ENEM por escola e os dados da educação básica 2015. Após o pré-processamento, o dataframe inclui as principais variáveis de interesse: as notas médias das cinco áreas do ENEM, a MEDIA\_GERAL calculada, e os indicadores criados relacionados à infraestrutura (INFRA\_BOOL\_SOMA, INFRA\_CRITICA\_SOMA, QT\_EQUIP\_TOTAL, QT\_FUNC\_TOTAL e FORMACAO\_DOCENTE). O dataset final está pronto para análises e modelagem, com todas as variáveis relevantes limpas, transformadas e organizadas.

## 3. Autocrítica

O grupo acredita que, até o momento, o trabalho está sendo conduzido de forma satisfatória, com boa aderência à metodologia CRISP-DM. Todas as etapas até a avaliação foram seguidas conforme o proposto, com organização e divisão de tarefas bem definidas. Do ponto de vista técnico, aprendemos a lidar com grandes volumes de dados educacionais, a criar variáveis relevantes e aplicar métodos estatísticos como a correlação. Do ponto de vista sociocomportamental, o trabalho em grupo exigiu organização, comunicação constante e colaboração para superar dúvidas e dificuldades técnicas.



Damos ao grupo a nota 9,0, pois cumprimos grande parte do escopo com qualidade, mas reconhecemos que ainda há espaço para melhorar a interpretação dos resultados e refinar a apresentação final.