PROJETO EM CIÊNCIA DE DADOS

SUMÁRIO

SEMESTRE	2025/1
PROJETO	A realidade paralela das instituições públicas: como dados do MEC apontam a proximidade entre os colégios militares e as redes privadas?
COMPONENTES DO GRUPO	Bernardo Possani Kirsch Bernardo Lykawka Medeiros Silva Luiza Flores Michel Iuri Cauan Capeletti Queiroz

Breve descrição do problema

A disparidade de desempenho entre instituições públicas e privadas no ENEM levanta dúvidas sobre a qualidade da educação pública. No entanto, colégios federais parecem fugir desse padrão, apresentando resultados comparáveis e superiores aos das escolas privadas. O problema é entender se esses colégios representam exceções ou um modelo viável de excelência dentro da rede pública.

Breve descrição da solução proposta

A proposta do grupo consiste em realizar uma análise exploratória e estatística dos microdados do ENEM, com foco no desempenho dos colégios federais em comparação com escolas públicas e privadas. O objetivo é identificar padrões, validar hipóteses e produzir idéias que orientem o debate sobre qualidade na educação pública.

Fases da Metodologia CRISP-DM

Fase	Tarefa principal	Conclusão
Compreensão do negócio	Definir problema e hipóteses	100%
Compreensão dos dados	Coleta dos microdados e identificação de colunas	80%
Preparação dos dados	Limpeza e filtragem das escolas	85%

Resumo do que foi concluído até o momento

Até o momento, o grupo já obteve os microdados do ENEM de 2015 e dados por escola entre 2005 e 2015. A análise exploratória inicial mostra que os colégios federais se destacam como exceções no cenário da educação pública. As dificuldades iniciais envolveram o volume dos dados e padronização dos nomes das instituições.

Autocrítica

O grupo avalia positivamente a aderência ao CRISP-DM até aqui, especialmente nas fases de compreensão e preparação dos dados. A equipe enfrentou dificuldades técnicas e de disponibilidade, mas superou com organização e divisão de tarefas. Nota atribuída: **9,5**. O grupo acredita que conseguirá entregar 100% do escopo, desde que mantenha o ritmo.

RELATÓRIO

1. Compreensão dos Dados

Coleta dos dados:

Os dados foram coletados diretamente do portal de dados abertos do INEP, utilizando os microdados do ENEM de 2015 e a base histórica por escola (2005 a 2015). O acesso foi feito via links oficiais, com arquivos em formato CSV e TXT, que exigiram tratamento para padronização de encoding e separadores.

Descrição dos dados:

Os microdados contêm informações detalhadas de desempenho por aluno (nota por área, tipo de escola, localização) e por escola (médias, número de inscritos, infraestrutura). As colunas mais relevantes incluem código da escola, rede de ensino, notas por área e UF

Análise exploratória dos dados:

Durante a exploração, foi aplicada uma segmentação por rede de ensino, mostrando que os colégios federais têm desempenho semelhante ao de escolas privadas. Utilizamos histogramas e boxplots para visualizar essas diferenças, além de calcular médias e desvios padrão

Verificação de qualidade dos dados:

Foi realizada a checagem de valores nulos, duplicações e registros inconsistentes (como notas zeradas em todas as disciplinas). Essas entradas foram filtradas para manter apenas dados relevantes e representativos.

2. Preparação dos Dados

Limpeza dos dados

Para a limpeza dos dados, foram adotados critérios específicos para remoção e a inclusão de variáveis com base na sua relevância para análise. As principais fases do processo envolveram a de filtragem geográfica e temporal no dataset de microdados por escola, que manteve apenas os dados que se referiam a partir do ano de 2009, já que a escala do ENEM antes era de 0-100, o que dificulta comparações com as edições mais atuais.

Além disso, foi realizado o tratamento de variáveis que tinham valores ausentes, como as colunas de média total e média das questões objetivas, que foram reconstruídas posteriormente com base nas notas disponíveis por área.

Para o conjunto de dados do ENEM de 2015, houve uma peneiração para grupos socioeconômicos e para a região Sul a fim de examinar se as respostas para as perguntas sobre o perfil socioeconômico de cada aluno das escolas federais se assemelham mais com as escolas privadas do que com as escolas da rede municipal e estadual.

Ademais, para os dados do ENEM de 2015, foi necessário filtrá-los para apenas as colunas que seriam utilizadas na pesquisa, já que o grande volume de dados seria impeditivo para os recursos do Kaggle processar. Em vista disso, foi criado um novo dataset apenas com os registros interessantes para a pesquisa.

Tendo em vista o novo dataset, 20 colunas apenas estão sendo utilizadas e elas são: SG_UF_ESC (uf da escola), TP_DEPENDENCIA_ADM_ESC (dependência administrativa (Escola)), NU_INSCRICAO (número da inscrição), TP_FAIXA_ETARIA (faixa etária do aluno), TP_ESCOLA (tipo de escola do Ensino Médio), IN_TREINEIRO (indica se o aluno era treineiro), TP_PRESENCA_CN (presença na prova de CN), TP_PRESENCA_CH, TP_PRESENCA_LC, TP_PRESENCA_MT, NU_NOTA_MT (nota da prova de matemática), NU_NOTA_LC, NU_NOTA_CH, NU_NOTA_CN, NU_NOTA_REDACAO, TP_STATUS_REDACAO (situação da redação do participante).

Criação de atributos e registros

Durante o processo de preparação dos dados, foi necessário a reconstrução dos campos de média total e média das questões objetivas com base na média das cinco áreas avaliadas (linguagens, ciências humanas, ciências da natureza, redação e matemática). Para o dataset dos dados dos alunos que realizaram o enem de 2015, foi realizado uma média simples das notas e incluído isso numa nova coluna chamada MEDIA_TOT.

Integração de dados

Para os dados do ENEM de 2015, foi realizada uma comparação entre os grupos socioeconômicos dos estudantes de cada tipo de escola (federal, municipal, estadual e privada), com o objetivo de verificar se os alunos das escolas federais apresentam perfil socioeconômico mais semelhante ao das escolas privadas do que ao das redes municipal e estadual. Para isso, será feita uma seleção das variáveis relacionadas à situação socioeconômica dos candidatos, com base nas respostas ao questionário socioeconômico do ENEM, que serão utilizadas na análise comparativa. Essa inspeção está sendo conduzida em conjunto com os dados de desempenho médio por escola, a fim de identificar possíveis similaridades entre os colégios federais e as instituições privadas, tanto em termos de resultados quanto de contexto socioeconômico dos alunos.

Descrição do dataset final

O dataset final contém informações provenientes dos microdados do ENEM, filtradas exclusivamente para escolas referentes ao período a partir de 2009. As principais variáveis incluem: unidade federativa (UF), médias de notas por área, média total, média das questões objetivas e o código da rede administrativa da escola. Além disso, foi incluída uma coluna binária para identificar colégios militares, com base na nomenclatura das instituições. Fora isso, será realizada uma pesquisa paralela e comparativa por rede de ensino a partir do tipo da escola e das variáveis socioeconômicas.

3. Autocrítica

Até o momento, o grupo teve uma boa aplicação da metodologia CRISP-DM, especialmente nas etapas de compreensão e preparação dos dados. O time enfrentou dificuldades de disponibilidade, mas superou com comunicação assertiva. Nota atribuída: 9,7. O grupo acredita que conseguirá entregar 100% do escopo, desde que mantenha o ritmo.