PROJETO EM CIÊNCIA DE DADOS

SUMÁRIO

SEMESTRE	2025/1
PROJETO	Presencial ou Remoto: Interfere no desempenho do aluno?
COMPONENTES DO GRUPO	Gustavo Sanford Bortolon Lucas Nory Ulson Victor dos Santos Cruz William de Oliveira Klein

Breve descrição do problema

Como estudantes, sempre nos questionamos sobre qual metodologia de ensino é mais eficiente: o modelo tradicional presencial ou o ensino a distância, que tem se popularizado entre os brasileiros que ingressam no ensino superior. Essa reflexão se torna ainda mais relevante quando observamos o contexto dos cursos da Licenciatura em Letras - Português e Espanhol, onde a modalidade EAD tem ganhado espaço. Diante disso, surge a dúvida: será que os alunos do ensino remoto apresentam desempenho acadêmico inferior aos do ensino presencial? E mais especificamente, como esses desempenhos se comparam quando avaliados por uma métrica nacional, como o ENADE 2021? Além disso, será que fatores como a renda familiar podem influenciar esses resultados?

Breve descrição da solução proposta

O projeto tem como objetivo responder à seguinte pergunta:

"Estudantes dos cursos EAD de Licenciatura em Letras - Português e Espanhol tiveram desempenho inferior aos estudantes de cursos presenciais no ENADE 2021?"

A proposta consiste em coletar, tratar e analisar os dados dos Microdados do ENADE 2021 e do Censo da Educação Superior 2021. As etapas incluíram:

Integração das duas bases de dados;

Análise do desempenho dos estudantes considerando a modalidade de ensino (EAD ou presencial);

Avaliação de como fatores socioeconômicos, como a faixa de renda, e fatores institucionais, como o tipo de instituição (pública ou privada) e a localização, podem impactar o desempenho dos alunos.

O projeto prevê a entrega de um dataset consolidado, análises estatísticas, visualizações (dashboards) e um relatório final respondendo à pergunta de pesquisa.

Fases da Metodologia CRISP-DM

Fase	Tarefas realizadas	Status (%)
1. Entendimento do Negócio	Definição da pergunta de pesquisa, alinhamento dos objetivos e entendimento dos impactos da análise.	**100%**
2. Entendimento dos Dados	Análise das duas bases, identificação das variáveis-chave, tipos de dados, qualidade e limitações dos dados.	**100%**
3. Preparação dos Dados	Limpeza dos dados, tratamento de valores ausentes, padronização de categorias, discretização de variáveis quando necessário e integração entre as bases.	**95%**
4. Modelagem	Planejada: testes estatísticos, análise de correlação, comparação de médias, desenvolvimento de dashboards.	**0%**
5. Avaliação	Será realizada após a modelagem para validar os insights e verificar se os resultados respondem à pergunta de pesquisa.	**0%**
6. Implementação	Desenvolvimento e entrega dos dashboards, relatório analítico e apresentação dos resultados.	**0%**

Resumo do que foi concluído até o momento

Até este ponto, foram concluídas as etapas de coleta, limpeza, preparação e integração dos dados do ENADE 2021 e do Censo da Educação Superior 2021, com foco nos cursos de Licenciatura em Letras - Português e Espanhol. Durante a análise exploratória, ficou evidente que não há dados de notas dos alunos de cursos EAD na área da computação, impossibilitando a comparação direta de desempenho entre modalidades no ENADE 2021.

Diante disso, a análise foi direcionada à Licenciatura em Letras - Português e Espanhol para o entendimento do perfil dos cursos e das instituições. Observou-se que a modalidade EAD possui um número significativamente maior de cursos e de alunos matriculados, especialmente em instituições privadas. Já os cursos presenciais apresentam uma distribuição mais equilibrada entre instituições públicas e privadas. Além disso, foi possível perceber que a oferta de cursos EAD está concentrada em determinadas regiões, enquanto os cursos presenciais possuem uma distribuição territorial mais ampla.

Portanto, o trabalho até aqui permitiu construir um panorama detalhado das características dos cursos de Licenciatura em Letras - Português e Espanhol nas modalidades EAD e presencial no Brasil, considerando quantidade de cursos, tipo de instituição e distribuição geográfica. A análise de desempenho permanece restrita aos alunos presenciais, devido à indisponibilidade de notas dos alunos EAD no ciclo do ENADE 2021.

Autocrítica

O grupo seguiu com boa aderência à metodologia CRISP-DM, principalmente nas etapas de entendimento, preparação e análise dos dados. No entanto, a descoberta da **ausência de notas dos alunos EAD no ENADE 2021** exigiu uma **reformulação dos objetivos**, tornando inviável a comparação de desempenho entre modalidades.

A análise, então, foi redirecionada para traçar o **perfil dos cursos e das instituições** nas modalidades presencial e EAD na área de Licenciatura em Letras - Português e Espanhol, considerando distribuição geográfica, tipo de instituição e número de matrículas.

Esse processo gerou importantes aprendizados, como a necessidade de validar a disponibilidade dos dados logo nas etapas iniciais, além do desenvolvimento de habilidades em tratamento de dados e resiliência para lidar com imprevistos.

O grupo atribui a si uma **nota 8,5**, reconhecendo o bom trabalho realizado, apesar do ajuste de escopo. Acreditamos que será possível **cumprir 100% do escopo no novo direcionamento**, com todas as etapas bem definidas e viáveis dentro do prazo.

-x-

RELATÓRIO

1. Compreensão dos Dados

Coleta dos dados

Os dados utilizados foram obtidos do portal oficial do **INEP (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira)**. Duas fontes principais foram utilizadas:

- **Microdados do ENADE 2021**, contendo informações detalhadas dos estudantes concluintes, como desempenho na prova (nota geral), modalidade do curso, e características socioeconômicas como faixa de renda.

- **Censo da Educação Superior 2021**, com informações sobre os cursos superiores ofertados no país, número de matrículas, tipo de instituição (pública ou privada), e localização.

Descrição dos dados

As principais bases e atributos utilizados foram:

Microdados ENADE 2021:

CO_MODALIDADE_ENSINO: modalidade do curso (1 = presencial, 0 = EAD)

NT_GER: nota geral do ENADE (principal métrica de desempenho)

QE_I08: faixa de renda familiar

CO_GRUPO: grupo de cursos, sendo 906 = Licenciatura em Letras - Português e Espanhol

CO_CURSO: código identificador do curso (chave de integração com o Censo)

Censo da Educação Superior 2021:

TP_MODALIDADE_ENSINO: modalidade do curso (confirmando a classificação do ENADE)

QT_MATRICULA_TOTAL: total de alunos matriculados por curso

TP_CATEGORIA_ADMINISTRATIVA: tipo da instituição (pública ou privada)

CO_IES e NO_UF_IES: código e estado da instituição

Essas informações permitiram contextualizar o ambiente educacional dos cursos analisados.

Análise exploratória dos dados

A exploração inicial dos dados foi realizada com o objetivo de compreender a estrutura geral das bases, o volume de registros e as variáveis disponíveis, permitindo organizar e planejar as transformações e filtragens necessárias. Essa etapa foi fundamental para garantir que os dados fossem adequados à pergunta de pesquisa e à construção do dataset final.

Foram utilizadas duas bases públicas disponibilizadas pelo INEP:

- **Microdados do ENADE 2021**: contém informações essenciais sobre os estudantes, como a nota geral no exame (NT_GER), utilizada como métrica de desempenho; a modalidade do curso (CO_MODALIDADE), que indica se é presencial ou EAD; além do código e grupo do curso (CO_CURSO e CO_GRUPO), que permitem filtrar os cursos desejados. Também foi utilizada a variável de faixa de renda familiar (QE_I08), que possibilita avaliar a influência da renda no desempenho dos estudantes.
- **Censo da Educação Superior 2021**: fornece dados institucionais sobre os cursos, como a modalidade (TP_MODALIDADE_ENSINO), o tipo de instituição (TP_CATEGORIA_ADMINISTRATIVA), o número total de matrículas (QT_MATRICULA_TOTAL) e o código do curso (CO_CURSO), que foi utilizado para integração com o ENADE. Também foram consideradas variáveis relacionadas à localização e à natureza da instituição, como o código da IES e as siglas do estado e município (CO_IES, NO_UF_IES, NO_MUNICIPIO_IES).

Inicialmente, a filtragem dos dados foi feita para selecionar os cursos do grupo 906, que corresponde à Licenciatura em Letras - Português e Espanhol. No entanto, após a análise exploratória, foi constatado que esse grupo não possuía registros suficientes na modalidade EAD, além de apresentar ausência total de notas (NT_GER) nessa modalidade, impossibilitando qualquer comparação de desempenho entre EAD e presencial.

Diante dessa limitação, foi realizada uma investigação sobre os demais grupos disponíveis nas bases, buscando aquele que apresentasse uma quantidade de registros mais equilibrada entre as modalidades presencial e EAD, além de garantir a presença de notas registradas. Após essa análise, o grupo 906 foi identificado como o mais adequado para a pesquisa, por apresentar uma diferença pequena entre as modalidades e quantidade suficiente de notas em ambas. Este grupo corresponde ao curso de Licenciatura em Letras - Português e Espanhol.

A partir da análise inicial dessas bases, foram definidos os seguintes objetivos para a exploração dos dados:

- Verificar a distribuição de alunos entre as modalidades presencial e EAD no curso de Licenciatura em Letras - Português e Espanhol (grupo 906);
- Observar a variação de desempenho entre essas modalidades, utilizando a nota geral do ENADE como métrica;
- Investigar a relação entre renda familiar e desempenho acadêmico, com base na variável QE_I08;
- Contextualizar os cursos com dados institucionais, como tipo de instituição (pública ou privada), utilizando informações do Censo;
- Identificar padrões iniciais, outliers e possíveis inconsistências nas variáveis de interesse.

Por conseguinte, para atingir os objetivos da exploração, foram aplicadas operações de leitura, concatenação e filtragem de dados utilizando a biblioteca Pandas. Os arquivos do ENADE foram unidos por concatenação horizontal (com base na ordem dos registros), garantindo o alinhamento das informações de curso, nota geral (NT_GER) e faixa de renda (QE_I08) de cada aluno. Em seguida, aplicou-se um filtro para manter apenas os registros do grupo 906.

Desenvolvendo, dessa forma, funções que permitem: calcular o total de registros, a distribuição de alunos por modalidade de ensino, a proporção de valores ausentes nas variáveis-chave e a média das notas por modalidade.

Entre os padrões identificados, observou-se que:

- O total de registros da base filtrada foi de 2.323;
- A distribuição de alunos ficou bastante equilibrada entre as modalidades, sendo 1.310 presenciais e 1.013 EAD;
- A maior parte dos estudantes possui renda familiar entre 3 a 4,5 salários mínimos (573) e entre 4,5 a 6 salários mínimos (567), indicando uma predominância de renda intermediária;
- A nota média dos alunos presenciais foi de 41,32, enquanto na modalidade EAD foi de 63,12;
- A proporção de valores ausentes foi de 24,75% na variável NT_GER e 12,91% na variável QE_I08;
- Especificamente na modalidade presencial, foram encontrados 953 registros com NT_GER presente e 357 ausentes;
- Na modalidade EAD, 795 registros possuem NT_GER presente e 218 estão ausentes.

Diante das análises realizadas, conclui-se que será necessário realizar o tratamento dos dados ausentes nas variáveis NT_GER e QE_I08, seja por meio da remoção de registros incompletos ou pela aplicação de estratégias de imputação, considerando os impactos sobre a qualidade da análise. Apesar de a base apresentar uma quantidade equilibrada de registros entre as modalidades presencial e EAD, foi identificado que ambas possuem proporções consideráveis de valores ausentes, especialmente na variável NT_GER, que impacta diretamente na análise de desempenho.

Por fim, é essencial garantir que os datasets estejam devidamente organizados e consistentes, considerando os tratamentos necessários, para que possam ser utilizados nas próximas etapas de análise, como na aplicação de métodos estatísticos mais robustos, modelagem preditiva ou na construção de visualizações que facilitem a interpretação dos resultados.

Verificação de qualidade dos dados

Durante a análise exploratória, foi possível avaliar os dados em questão ao escopo da nossa questão de pesquisa. Além de verificar possíveis inconsistências, podemos observar os valores ausentes e limitações que poderiam impactar a integridade das análises.

A validação consistiu na busca de dados ausentes nas principais variáveis que irão suprir nossos objetivos (nota geral e renda familiar). Ademais, na coerência dos dados categóricos, como a modalidade de ensino e faixa de renda. Como resultado, observou-se que NT_GER apresenta 24,75% de valores ausentes e QE_I08, 12,91%, o que exige atenção nas etapas seguintes.

Diferente do grupo anterior, no grupo 906 há notas registradas tanto na modalidade presencial quanto na EAD, viabilizando a análise comparativa. Também foi confirmada a consistência dos dados categóricos e a ausência de registros duplicados ou erros estruturais.

Portanto, embora os dados estejam bem estruturados, a presença de valores ausentes e a necessidade de balanceamento entre as modalidades são os principais pontos que devem ser tratados na preparação dos dados.

2. Preparação dos Dados

Nesta seção, foram descritas as atividades realizadas para a construção do dataset final, as atividades incluem limpeza, criação de atributos, inserção de registros, integração de bases etc. E ao final, temos uma descrição do estado do dataset que será utilizado para a modelagem deve ser realizada.

Limpeza dos dados

Apresenta o raciocínio para se incluir ou remover features no dataset final, bem como o que foi realizado com dados faltantes, quais transformações foram realizadas etc.

A limpeza dos dados teve como foco principal garantir a qualidade e consistência das variáveis essenciais para a análise, com base nos seguintes critérios:

- Filtragem por curso: inicialmente foi considerado o grupo `CO_GRUPO = 906` (Licenciatura em Letras - Português e Espanhol), mas, após constatação de ausência de notas no EAD, foi adotado o grupo `906` (Licenciatura em Letras – Português e Espanhol), que apresentou maior equilíbrio entre as modalidades presencial e EAD, com notas registradas em ambas.

- Remoção de valores ausentes: registros com dados ausentes nas variáveis NT_GER (nota geral do ENADE) e QE_I08 (faixa de renda familiar) foram removidos, por serem variáveis centrais nas análises de desempenho e perfil socioeconômico.

- Remoção de valores inválidos: também foram excluídos os registros em que `NT_GER` era igual a 0.0, mesmo não sendo nulos, pois representam alunos que não participaram da avaliação e poderiam distorcer os resultados.

- Remoção da coluna `QE_I08`: após seu mapeamento para `FAIXA_RENDA`, a coluna `QE_I08` foi removida do dataset final para evitar redundância e manter apenas variáveis legíveis.

- Validação de categorias: foi garantido que os códigos das modalidades (`0` e `1`) e faixas de renda (A a G) estivessem dentro dos padrões definidos pelo INEP.

Criação de atributos e registros

Com o objetivo de facilitar a análise e interpretação dos dados, foram criados os seguintes atributos derivados:

- MODALIDADE: criado a partir de `CO_MODALIDADE`, convertendo os valores numéricos em categorias legíveis: `"Presencial"` ou `"EAD"`.

- FAIXA_RENDA: resultado do mapeamento da variável `QE_I08`, traduzindo os códigos literais (A–G) para descrições completas dos intervalos salariais correspondentes, facilitando a leitura e a análise por perfil socioeconômico.

- NO_UF_IES: nome do estado da Instituição de Ensino Superior, incorporado a partir da base do Censo de Instituições por meio da chave `CO_IES`.

Nenhum novo registro foi criado manualmente. Todas as observações do dataset final correspondem a estudantes reais extraídos dos microdados oficiais.

Integração de dados

Foram utilizadas duas bases públicas disponibilizadas pelo INEP:

- Microdados do ENADE 2021: com informações individuais dos estudantes (notas, curso, renda, modalidade);
- Censo da Educação Superior 2021: com informações institucionais das IES (tipo, localização, natureza jurídica).

A integração foi feita por meio das chaves primárias `CO_CURSO` e `CO_IES`, permitindo associar dados dos cursos e instituições aos dados dos alunos.

Durante a integração, foi observado que diversas colunas se repetem entre os arquivos, com dados institucionais ou geográficos idênticos (UF, município, região). Para lidar com essa redundância, foram aplicadas as seguintes estratégias:

1. Padronização de nomes e códigos para garantir o sucesso das junções (`merge`);
2. Identificação e remoção de colunas duplicadas após o merge, mantendo apenas os campos relevantes;
3. Seleção preferencial de colunas mais específicas e completas (ex: priorização de UF do IES em vez da UF do curso).

Feature redundantes

Nome da coluna	Dado armazenado
NU_ANO_CENSO	Ano de referência do Censo da Educação Superior
CO_IES	Código único da Instituição de Ensino Superior
TP_ORGANIZACAO_ACADEMICA	Tipo de organização acadêmica da IES (pública ou privada)
TP_CATEGORIA_ADMINISTRATIVA	Categoria administrativa da IES

CO_UF_CURSO / CO_UF_IES	Unidade da Federação
CO_MUNIC_CURSO / CO_MUNICIPIO_IES	Município onde o curso ou a IES está localizada
CO_REGIAO_CURSO / CO_REGIAO_IES	Região geográfica

Descrição do dataset final

Etapas Realizadas:

1. Filtragem da Área de Conhecimento

- Selecionamos apenas o curso Licenciatura em Letras - Português e Espanhol com base no código de grupo (`CO_GRUPO = 906`).

2. Integração dos Dados

- Os microdados do ENADE (que contêm informações de desempenho dos estudantes) foram integrados com os dados do Censo Superior (instituições e cursos).

3. Tratamento de Redundâncias

- Remoção de colunas duplicadas ou sobrepostas, mantendo apenas versões padronizadas (ex: mantida apenas uma versão de `CO_UF`, `CO_MUNICIPIO`, etc.).
- Consolidamos informações institucionais (como UF, município, tipo de instituição) em nível de curso.

4. Limpeza dos Dados

- Remoção de registros com **notas inválidas ou ausentes** no ENADE.
- Conversão de variáveis para tipos apropriados (ex: categóricas e numéricas).
- Padronização dos nomes de regiões, estados e instituições.

5. Criação de Novas Colunas

Coluna criada	Descrição
MODALIDADE	Modalidade do curso (Presencial ou EAD)
CO_Curso	Código identificador do curso
NO_UF_IES	Estado da instituição de ensino
NT_GER	Nota geral média dos alunos da IES no curso, conforme dados consolidados do ENADE
QE_I08	Código da faixa de renda familiar do estudante
FAIXA_RENDA	Descrição da faixa de renda
TP_CATEGORIA_ADMINISTRATIVA	Tipo de instituição (1 = privado, 0 = pública)

3. Autocrítica

Até o momento, o grupo acredita ter feito um bom progresso no desenvolvimento do trabalho, mantendo uma boa organização e aderência às etapas da metodologia **CRISP-DM**. Conseguimos compreender claramente o problema de negócio — analisar comparativamente o desempenho dos cursos Licenciatura em Letras - Português e Espanhol no ENADE 2021 entre as modalidades presencial e remota — e, a partir disso, fizemos uma coleta e integração de dados consistente com os objetivos definidos.

A etapa de **Entendimento dos Dados** foi bem explorada, e a fase de **Preparação dos Dados** está em estágio avançado. Identificamos e tratamos dados redundantes, realizamos limpezas relevantes e já planejamos a criação de variáveis úteis para a etapa de modelagem, como a categorização dos cursos por modalidade e o agrupamento por características institucionais. Ainda não concluímos a **Modelagem**, mas estamos nos preparando tecnicamente para executá-la com qualidade e embasamento.

O grupo tem trabalhado de forma colaborativa, com boa divisão de tarefas, e houve evolução tanto em aspectos técnicos (como limpeza e análise de dados em Python/Pandas) quanto em aspectos comportamentais, como comunicação, organização e tomada de decisões em grupo. Aprendemos também a lidar com dados reais e complexos, o que exigiu pensamento crítico e flexibilidade.

Atribuímos a nota **9,0** ao nosso trabalho até o momento. Consideramos que o grupo demonstrou um bom entendimento do problema, planejamento eficaz e progresso significativo nas etapas iniciais do CRISP-DM.

Apesar de ainda não termos concluído todas as fases do projeto, acreditamos que estamos no caminho certo e com base sólida para finalizar o trabalho com qualidade.