PROJETO EM CIÊNCIA DE DADOS

SUMÁRIO

SEMESTRE	2025/1
PROJETO	Nome do projeto/Empresa
COMPONENTES DO GRUPO	Alexya Tatiana Silva Ungaratti Filipe Tesche Roman Lorenzo Schavinski Eulalio de Souza Rennan Wictor Vieira Nascimento Carpes dos Santos

Breve descrição do problema

No contexto da educação de nível superior do Brasil, desejamos analisar o desempenho das Universidades no que diz respeito aos cursos da área da saúde, como Medicina, Nutrição, Enfermagem e afins. Nosso objetivo tange analisar influências institucionais, mapear a distribuição regional e identificar as instituições de destaque e lacunas.

Breve descrição da solução proposta

Propomos uma análise estatística e exploratória dos microdados do ENADE, combinando-os com outras bases de dados do MEC referentes às instituições de Ensino Superior. A ideia é filtrar os cursos da área de saúde e correlacionar as notas com as características das IES – como porte, modalidade e tipo, além de analisar a região onde cada uma delas se encontra.
Entregas pretendidas:

1. Seleção e coleta de dados;
2. Análise dos dados;
3. Relatório técnico;
4. Apresentação final.

Fases da Metodologia CRISP-DM

1	Entendimento de dados	Coleta e exploração inicial dos datasets e elaboração da pergunta do projeto.	100%
2	Preparação de dados	Limpeza, junção, filtragem dos dados e definição de variáveis.	80%
3	Modelagem	Escolha de métodos estatísticos e técnicas exploratórias.	50%
4	Avaliação	Interpretação dos resultados e validação das	50%

		análises com base nos objetivos.	

Resumo do que foi concluído até o momento

Etapa	Fase	Desafios/Superações	Status
1	Primeira Entrega Parcial – *Data Understanding*	O grupo teve desafios em chegar a um consenso da pergunta do projeto, pois havia inúmeras alternativas, dentro do escopo sugerido. Conseguimos, após uma análise profunda dos Datasets, chegar a uma definição e avançar no nosso projeto.	Concluído
2	Segunda Etapa Parcial – *Data Preparation/Modeling*	Apesar da riqueza do dataset unificado, uma avaliação inicial da qualidade dos dados brutos (antes das etapas de limpeza e transformação mais profundas) revelou desafios significativos. A análise preliminar de valores ausentes indicou que 351 das 361 colunas continham algum nível de dados faltantes, algumas com uma porcentagem considerável de ausência. Adicionalmente, observou-se a necessidade de padronizar representações de dados categóricos (que estavam em formato de código numérico ou textual) para facilitar a interpretação. Também foram identificados	Em andamento

		outliers pontuais, como em idades de participantes, e colunas específicas de questionários aplicados apenas a um subconjunto de cursos (licenciaturas) que resultavam em alta incidência de valores não preenchidos para os demais. Essas constatações justificaram a aplicação de um conjunto de etapas de pré-processamento, detalhadas a seguir, com o objetivo de preparar um dataset mais robusto e confiável para a análise exploratória.	
3	Entrega Final	Nesta etapa, pretendemos realizar a análise exploratória dos dados, chegando à resposta para a nossa pergunta.	Não iniciado

Autocrítica

Até o momento, demonstramos boa aderência à metodologia CRISP-DM, especialmente na fase inicial do projeto.
Nesta segunda etapa, destacamos a importância da boa compreensão do enunciado e a comunicação efetiva entre os membros do grupo.
Nota atribuída ao grupo: 7,5.
Justificativa: Acreditamos que o grupo ainda possa evoluir a questão técnica, utilizando as metodologias aprendidas em aula.
Desde que o grupo mantenha o foco e a organização, bem como a organização na distribuição de tarefas, a expectativa de cumprimento do projeto é de 100% do escopo pretendido.

-x-

RELATÓRIO

1. Compreensão dos Dados

Coleta dos dados

Os dados foram coletados de três fontes oficiais do Ministério da Educação (MEC), disponíveis publicamente:

1. Microdados ENADE 2021: 43 arquivos .txt contendo variáveis relacionadas ao desempenho e perfil dos estudantes.
2. Cadastro de Cursos Superiores (2023): arquivo .CSV com dados administrativos e descritivos dos cursos superiores no Brasil.
3. Cadastro de Instituições de Ensino Superior (IES - 2023): arquivo .CSV com dados institucionais das IES brasileiras.

Esses arquivos foram lidos e unificados utilizando a biblioteca pandas no Python. O processo envolveu carregamento com definição explícita de tipos (dtypes), remoção de duplicatas e validações das chaves de unificação.

Descrição dos dados

Cada fonte de dados contém diferentes blocos de informações:

1. ENADE 2021 (43 arquivos): informações acadêmicas, notas, respostas do questionário do estudante e dados institucionais.
    a. Principais campos: NU_ANO, CO_CURSO, NT_GER, NT_FG, TP_SEXO, NU_IDADE, QE_I01 a QE_I92, entre outros.
2. Cadastro de Cursos: 671.610 registros, 202 colunas.
    a. Chaves: CO_CURSO, NU_ANO.
3. Cadastro de IES: 2.580 registros, 84 colunas.
    a. Chave principal: CO_IES.

O dataset final consolidado (df_final_completo) apresenta 112.061 registros e 361 colunas.

Análise exploratória dos dados

A exploração inicial foi feita com:

1. .info(): para tipos de dados e colunas.
2. .describe(): para estatísticas descritivas de colunas numéricas.
3. .value_counts(): para variáveis categóricas, como sexo (TP_SEXO) e renda (QE_I08).
4. Identificação de colunas com maior número de valores faltantes.
5. Mapeamento de códigos categóricos para valores descritivos (ex: CO_CATEGAD → “Pública Federal”).

Observou-se uma alta diversidade de tipos de dados, uma quantidade elevada de dados faltantes (351 colunas com algum NaN), presença de outliers (idades menores do que 15 anos) e colunas irrelevantes.

Verificação de qualidade dos dados

Foram utilizados os seguintes critérios:

1. Proporção de valores nulos por coluna;
2. Detecção de inconsistências como idades irreais;
3. Presença de colunas com preenchimento seletivo;
4. Verificação de duplicatas por chaves NU_ANO e CO_CURSO.

Conclusão: foi necessária aplicação de múltiplas estratégias de pré-processamento para garantir a consistência dos datasets.

2. Preparação dos Dados

Limpeza dos dados

Foram aplicadas as seguintes ações:

1. Remoção de colunas com preenchimento parcial e baixa relevância (13 colunas do questionário específicas para licenciaturas).
2. Imputação de dados faltantes:
    a. Numéricos: mediana.
    b. Categóricos: moda.
    c. Genérico:
        i. object: "Não Informado".
        ii. number: 0.
3. Tratamento de outliers:
    a. Idades menores que 15 anos foram substituídas por NaN.

Criação de atributos e registros

Foram criadas colunas descritivas mapeando códigos:

1. CO_CATEGAD_DESC – Categoria administrativa da IES.
2. QE_I08_DESC – Faixa de renda familiar.
3. QE_I89_DESC a QE_I92_DESC – Opiniões em formato textual.

Integração de dados

1. ENADE unificado a partir de 43 arquivos por NU_ANO e CO_CURSO.
2. Cadastro de Cursos unido via CO_CURSO e NU_ANO.
3. IES unido via CO_IES.

Colunas com nomes semelhantes foram diferenciadas usando sufixos _enade, _curso_mec, _ies_mec.

Descrição do dataset final

O dataset final consolidado, df_final_completo, após todas as etapas de integração, apresenta as dimensões de **112.061 observações (linhas)** e **361 variáveis (colunas)**. Este é o conjunto de dados que servirá de base para as subsequentes etapas de análise exploratória, avaliação detalhada da qualidade dos dados e o processo de pré-processamento descrito neste relatório.

3. Modelagem

Técnicas e suposições de modelagem

Ainda em fase de preparação para testes, supõe-se que os dados sejam utilizados em modelos de classificação (tipo de instituição, curso, desempenho). Além disso, tem-se como pré-requisitos garantidos os dados limpos e estruturados, sem valores ausentes e atributos genéricos mapeados.

Projeto de testes e experimentos

Descrição dos modelos

Avaliação dos modelos