Apresentação de Limpeza e Pré-processamento de Dados

Esta apresentação detalha o processo de limpeza e pré-processamento de dados realizado para o projeto, abordando desde a imputação de valores faltantes até a engenharia de features, garantindo que os dados estejam prontos para a modelagem.

Membros: Victor Matheus (01716714),
José Humberto Silva de Araújo – 01589405,
Naeliton Chavez - 01594737

Limpeza e Imputação de Dados: Garantindo a Qualidade

Nesta seção, abordamos as etapas cruciais de limpeza e imputação para preparar os dados de forma segura e eficaz. Detalhamos as ações tomadas para tratar valores faltantes, duplicatas e outliers, garantindo a integridade dos dados para a modelagem.

Faltantes Numéricos

Identificados em 4 colunas, utilizamos a imputação por **Mediana** via Pipeline. A mediana é robusta contra outliers e assimetrias, preservando a distribuição.

Faltantes Categóricos

Para colunas categóricas, aplicamos a imputação por **Moda** (categoria mais frequente), uma estratégia padrão que mantém a distribuição original dos dados.

Duplicatas

Oito linhas idênticas foram **removidas** para garantir a integridade do conjunto de dados e evitar o viés no treinamento do modelo.

Outliers e Data Leakage

Detectamos aproximadamente 80 outliers, mas **nenhum foi removido**, pois eram valores legítimos de desempenho. Seu impacto será mitigado pelo Standard Scaling. Para evitar Data Leakage, todas as transformações foram ajustadas **APENAS** no conjunto de **TREINO**, utilizando Pipelines do Scikit-learn para consistência.

Transformações e Encoding: Preparando Variáveis

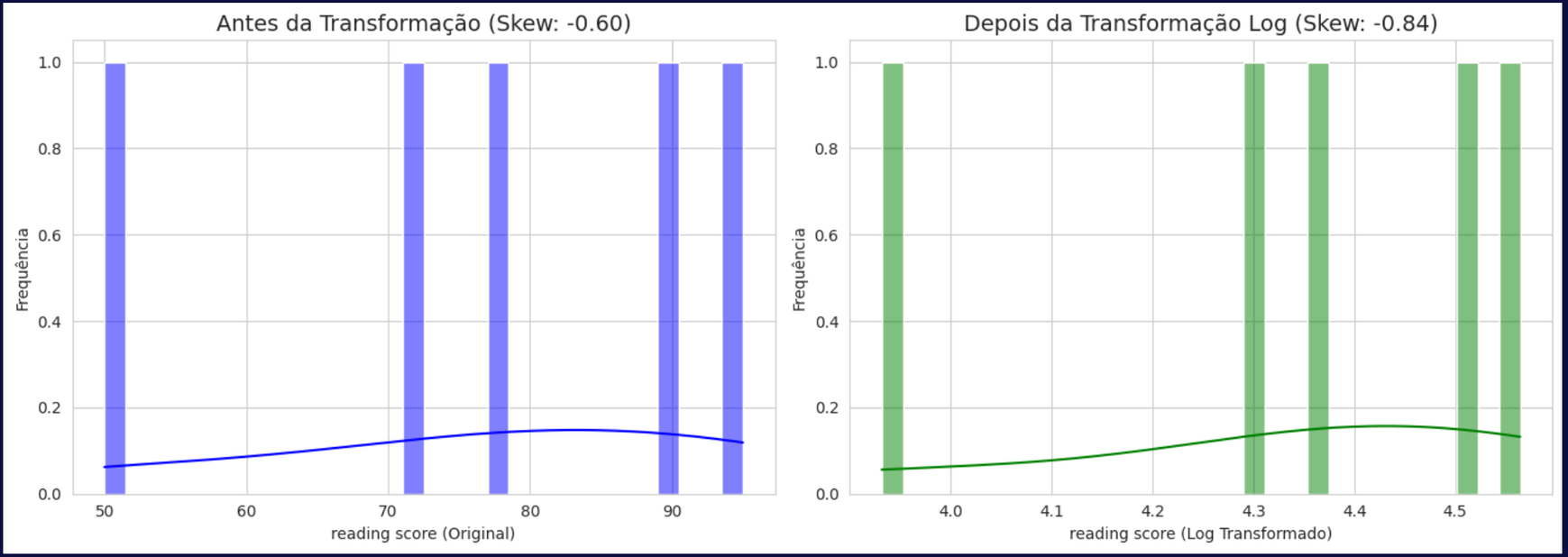
Nesta fase, ajustamos a distribuição de variáveis numéricas e codificamos categóricas, otimizando-as para algoritmos de Machine Learning.

Ajuste de Distribuição

A coluna `study_hours_week` (assimétrica positiva) recebeu uma **Transformação Logarítmica (np.log1p)**, aproximando-a de uma distribuição normal e reduzindo o peso da cauda longa. A `final_grade` foi mantida, pois sua assimetria é aceitável e a escala original é prioritária.

Encoding de Variáveis Categóricas

- **Nominais:** Utilizamos **One-Hot Encoding** com `drop_first=True`, resultando em 9 novas colunas. Essa abordagem evita a multicolinearidade em modelos lineares.
- **Ordinais:** O **Label Encoder** foi aplicado à coluna `parental_education`, preservando a ordem lógica das categorias (e.g., low < medium < high).



Estas transformações garantem que os dados estejam no formato ideal para os modelos, melhorando o desempenho e a interpretabilidade.

Dimensões Finais do Dataset

8

Colunas Originais

O dataset original possuía 8 colunas de dados.

O número de linhas permaneceu o mesmo, mas a complexidade e a qualidade das features foram significativamente aprimoradas.

16

Colunas Finais

Após One-Hot Encoding e Feature Engineering, o número de colunas aumentou para 16.

Cada nova coluna representa uma característica mais detalhada ou uma transformação numérica essencial para o aprendizado de máquina.

Feature Engineering: Criando Poder Preditivo

Introduzimos duas novas features cuidadosamente desenvolvidas para capturar insights que as colunas originais não ofereciam, enriquecendo o poder preditivo do nosso modelo.

study_intensity

Calculada como `study_hours_week / sleep_hours`, esta razão mede a prioridade do estudo sobre o descanso. Esperamos uma **alta correlação positiva** com a `final_grade`, refletindo o esforço e sacrifício dedicados aos estudos.

high_parental_support

Esta é uma variável binária que combina `(parental_education Alto) E (tutoring Yes)`. Ela cria um indicador robusto de **ambiente acadêmico com alto suporte**, prevendo-se que seja um preditor mais forte do que as variáveis isoladas.

Essas novas features são cruciais para aprimorar a compreensão do impacto do ambiente e hábitos de estudo no desempenho acadêmico.

Dados Prontos para Modelagem e Scaling

Concluímos a fase de pré-processamento, e os dados estão agora limpos, transformados e padronizados, prontos para a próxima etapa de modelagem e avaliação.

- **Notebook:** O processo completo está documentado em `notebooks/02_Preprocessamento.ipynb`.
- **Dataset Limpo:** Disponível em `data/students_clean.csv`.
- **Scaler Salvo:** O `scaler.pkl` foi salvo em `models/` para uso futuro.

Scaling (Q11)

Aplicamos o **StandardScaler** em 6 features numéricas (incluindo `study_intensity`). Escolhemos o StandardScaler (média 0, desvio-padrão 1) por sua adequação à maioria dos algoritmos de Machine Learning, garantindo que todas as features contribuam igualmente.

Por que Salvar o Scaler? (Q12)

Salvar o scaler garante que, durante o teste ou em produção, qualquer novo dado será transformado usando **EXATAMENTE** as mesmas estatísticas de média e desvio-padrão do conjunto de treino, mantendo a **CONSISTÊNCIA** e a integridade do modelo.

Dimensões Finais

Linhas	2510	2502
Colunas	14	16

Próximo Passo:

Os dados estão agora totalmente preparados. A **Etapa 3: Modelagem e Avaliação** pode começar!