PProductions

Análise Estratégica Cinematográfica - Vitória Freire

Relatório executivo com hipóteses testadas, métricas e plano de modelagem.

SUMÁRIO EXECUTIVO

DADOS:
- 999 filmes de 1920 a 2020
- 42 variáveis analisadas
- Cobertura temporal: 100 anos

DESTAQUES:
- Nota média IMDb: 7.95
- Correlação (log) receita-popularidade: 0.568
- Gênero com maior receita média: ver seção de receita
- Duração típica do top 20%: ~120-140 min
- Sentimento predominante nas sinopses: Neutro

RECOMENDAÇÃO PRINCIPAL:
Focar nos perfis com maior score na matriz de decisão: {'Drama (120min)': np.float64(7.425), 'Ação (130min)': np.float64(8.475), 'Comédia (100min)': np.float64(6.8)}

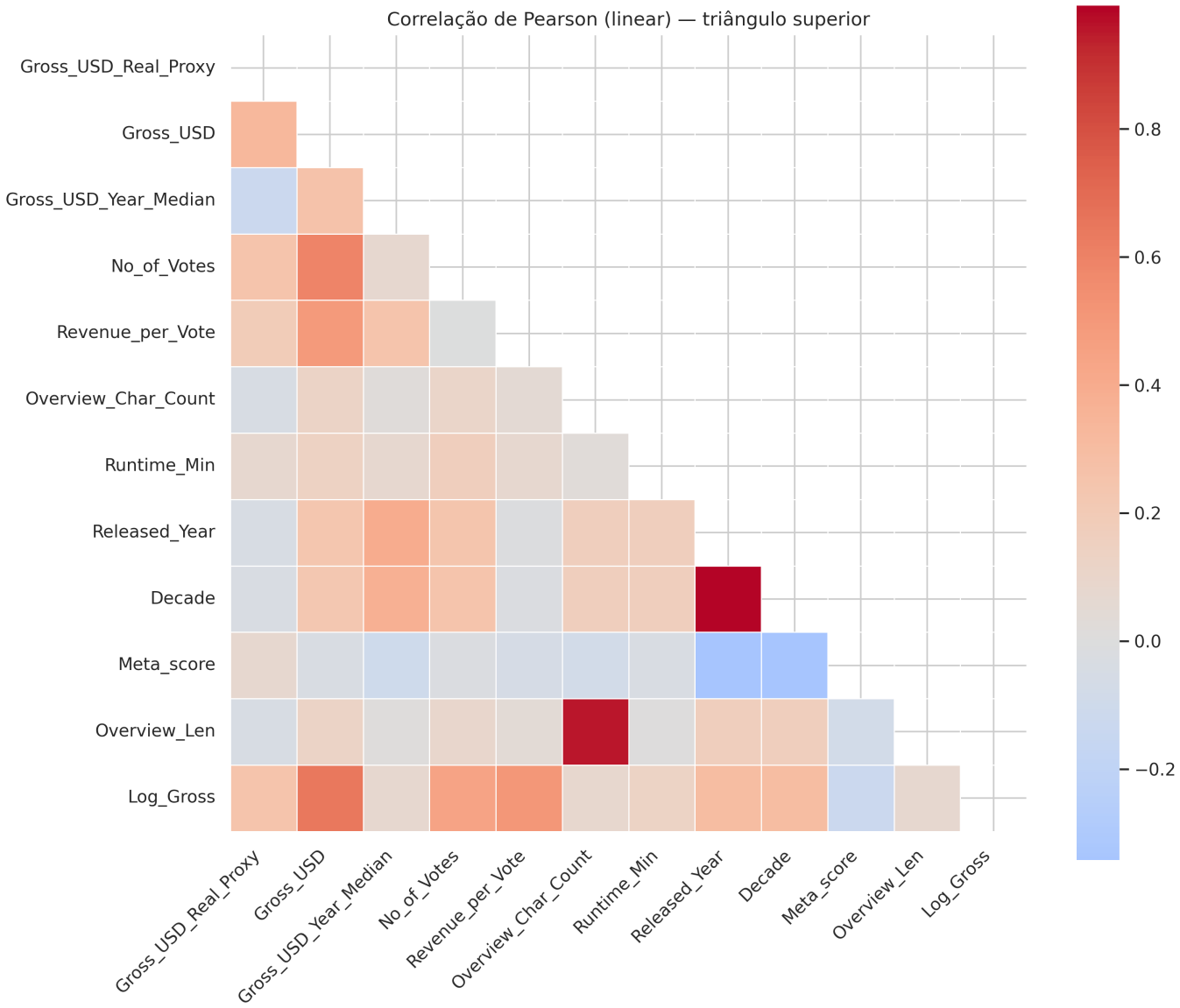
PLANO DE MODELAGEM (Pré-execução)

Problema: regressão (prever IMDB_Rating).
Métricas: MAE (principal), RMSE e R2 (comparação). Baseline: mediana do treino.
Features: numéricas (Meta_score, No_of_Votes + log1p, Votes_Z_By_Year, Runtime_Min, Released_Year, Overview_Len, Overview_Polarity, Overview_Subjectivity, Genre_Count, Revenue_per_Vote, Gross_USD_Real_Proxy), categóricas (Genre_Primary, Certificate, Cinema_Era, Runtime_Category, Overview_Sentiment_PT, Performance_Segment), texto (TF-IDF/SVD de Overview).
Validação: K-Fold estratificada temporalmente por ano ou GroupKFold por diretor.
Interpretação: SHAP/Permutation Importance no modelo final.
Vazamento: evitar estatísticas globais; usar encoding por fold. Hold-out: 'The Shawshank Redemption' fora do treino.

HIPÓTESES E RESULTADOS

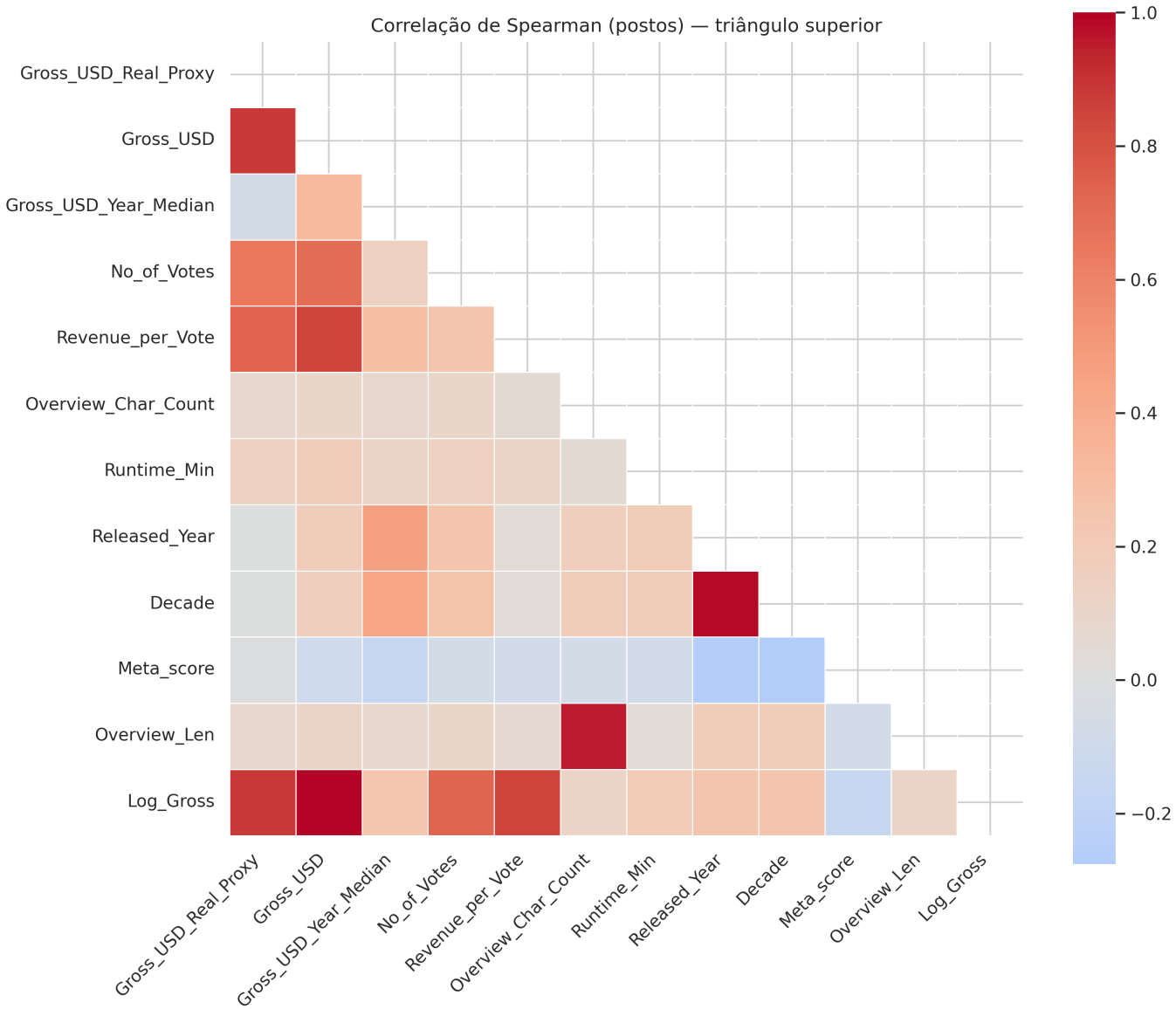
- Rating ~ Gênero (Top5): ANOVA F=2.843, p=0.0233, eta2=0.014 | Kruskal H=7.920, p=0.0945, ε2=0.005
- Rating ~ Polaridade da sinopse: Spearman ρ=-0.042, p=0.1828 |
- Rating ~ Categoria de duração: ANOVA F=22.590, p=0.0000, eta2=0.064 | Kruskal H=53.140, p=0.0000, ε2=0.050

Corr Pearson Upper



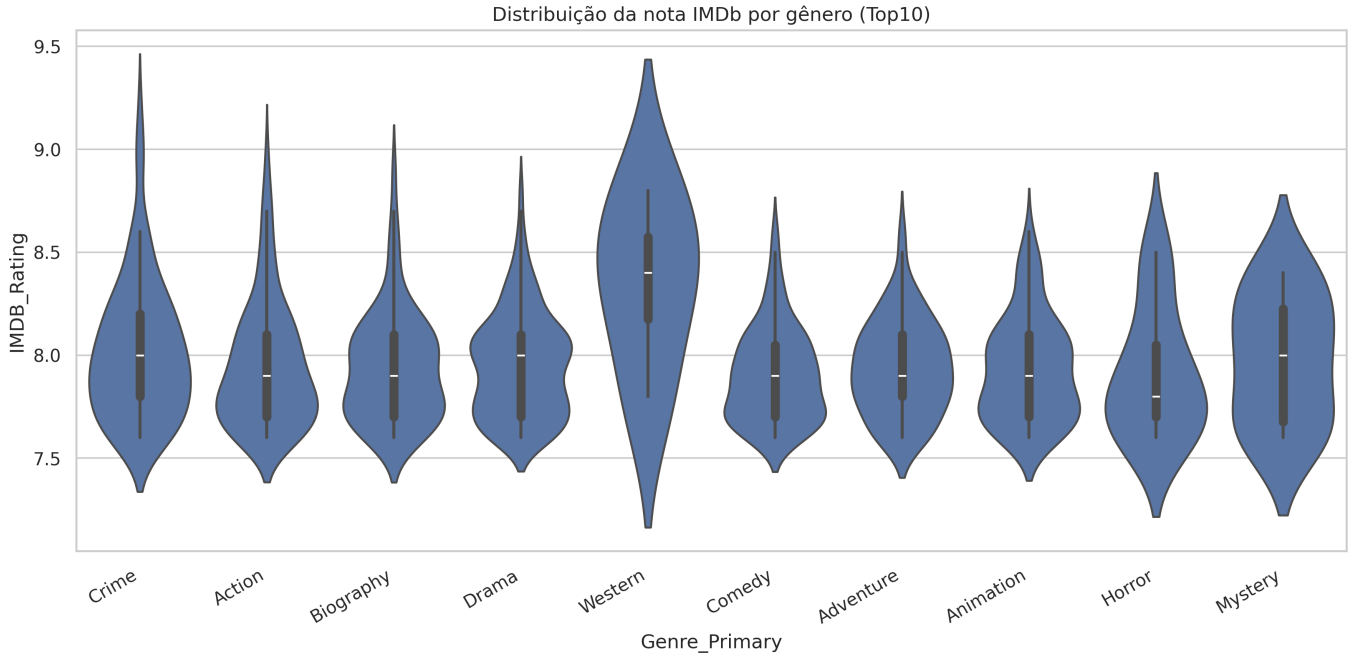
Insight: Correlação linear entre variáveis numéricas.

Corr Spearman Upper



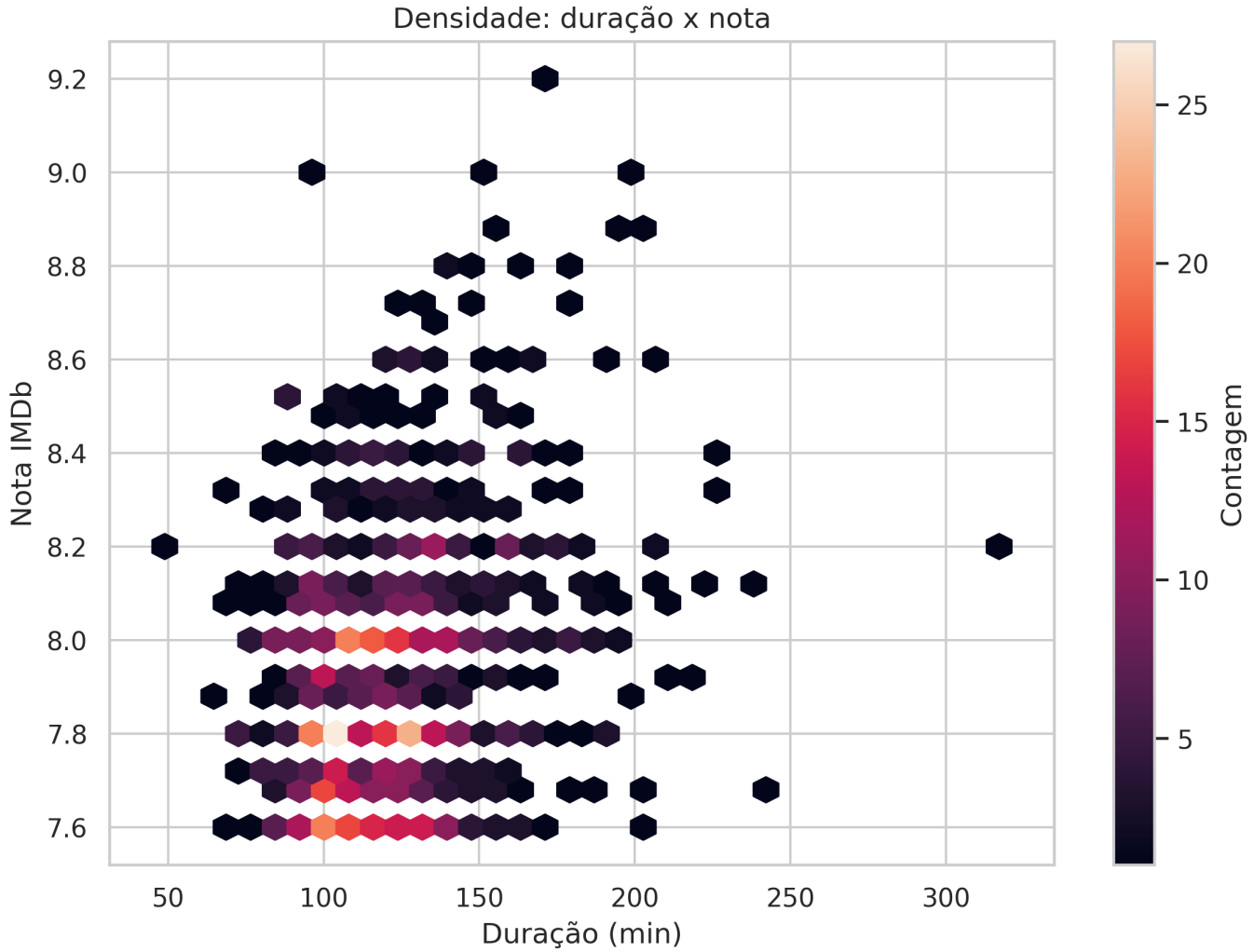
Insight: Correlação por rankings (mais robusta a outliers).

Violin Rating By Genre



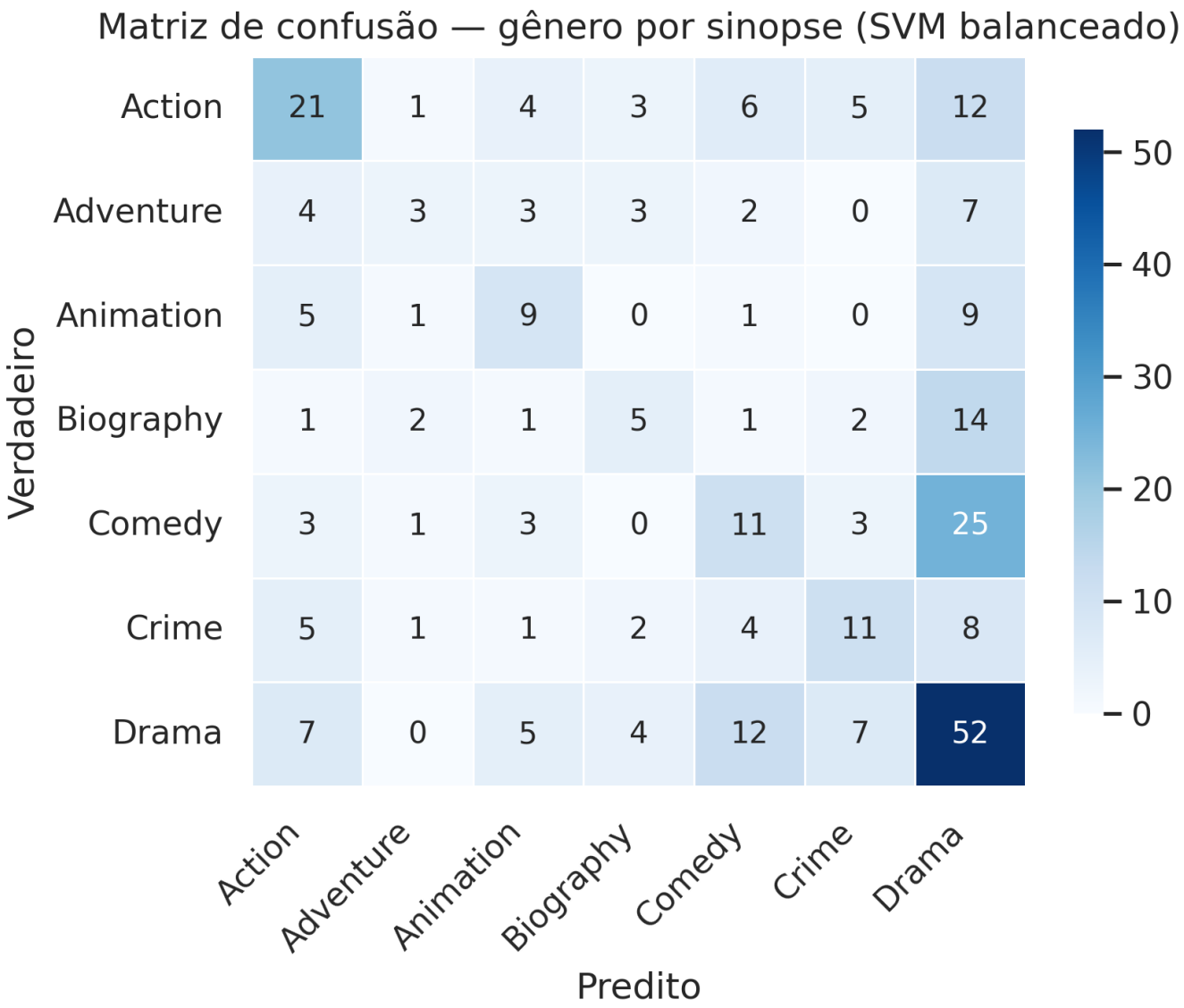
Insight: Forma e dispersão da nota por gênero (Top10).

Hexbin Runtime Rating



Insight: Densidade na relação duração × nota.

Cm Overview Genre



Insight: Diagonal = acertos; off-diagonal = confusões entre gêneros.

Reprodutibilidade

- Seeds fixos (Python/Numpy): 42
- Artefatos: figs/ e artifacts/ (CSV processado)
- Ambiente: Google Colab (Linux), PDF via FPDF
- Data de geração: 03/09/2025