Relatório de Modelagem e Validação

Objetivo: Construir e avaliar modelos preditivos

Dataset: processed_data.csv

Autor: Flávia Souza e Vinicius Loeblein

1. Introdução

Este relatório apresenta a modelagem preditiva do dataset processado, com avaliação, comparação e interpretação dos modelos testados.

2. Visualização Inicial dos Dados

2.1 Amostra dos Dados (3 primeiras linhas)

age	gender	job_type	daily_social_media_time	social_platform_preference	number_of_notifications	wo
56	Male	Unemployed	4.180940	Facebook	61	
46	Male	Health	3.249603	Twitter	59	
32	Male	Finance	3.113418	Twitter	57	

3. Modelagem

3.1 Modelo Baseline (Média)

O modelo baseline prevê sempre a média dos valores de produtividade observados no conjunto de treino.

Métrica	Valor
MAE	1.49
R^2	-0.00

3.2 Regressão Linear

A Regressão Linear é um modelo estatístico que tenta ajustar uma relação linear entre as variáveis preditoras e a variável alvo.

Métrica	Valor
MAE	1.49
R^2	-0.00

3.3 Random Forest

Random Forest é um modelo de ensemble que combina múltiplas árvores de decisão para melhorar a precisão e evitar overfitting.

Métrica	Valor
MAE	1.49
R^2	-0.00

3.4 XGBoost

XGBoost é um algoritmo de boosting eficiente e poderoso que combina múltiplos modelos fracos para criar um forte modelo preditivo.

Métrica	Valor
MAE	1.50
R^2	-0.01

4. Comparação de Modelos

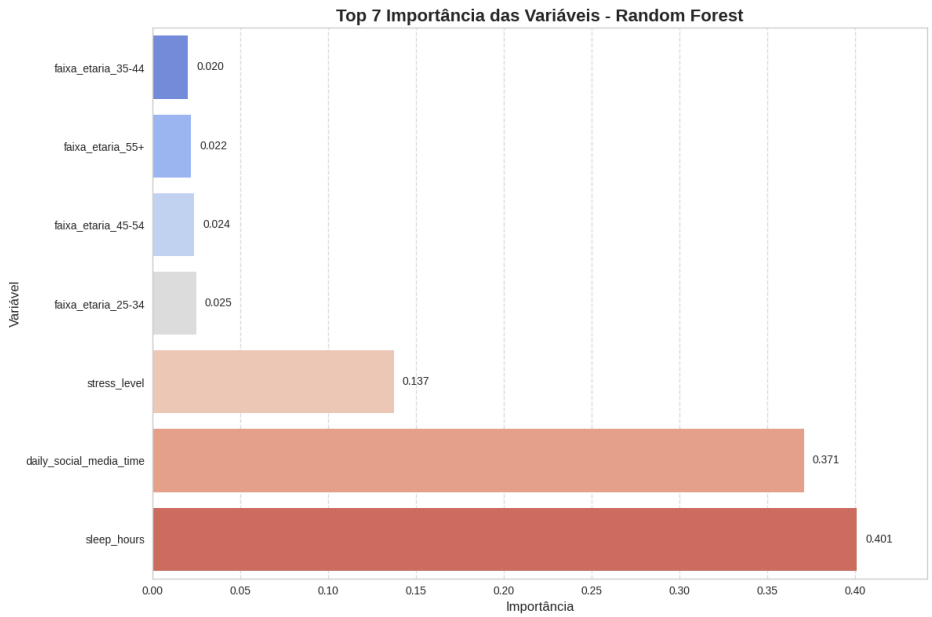
A tabela abaixo mostra a comparação dos modelos ordenada pelo MAE (menor é melhor):

Modelo	MAE	R^2
Baseline (Média)	1.49	-0.00
Regressão Linear	1.49	-0.00
Random Forest	1.49	-0.00
XGBoost	1.50	-0.01

5. Interpretação do Melhor Modelo

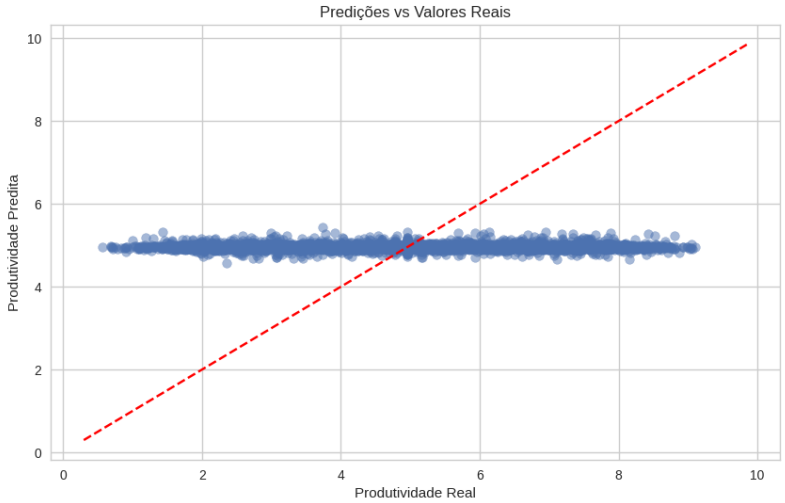
O modelo Random Forest foi identificado como o melhor modelo com base nas métricas de avaliação. Para entender melhor o impacto das variáveis na predição, apresentamos abaixo o gráfico das **Top 7 variáveis mais importantes** segundo o modelo Random Forest.

Este gráfico mostra quais features mais influenciam a capacidade do modelo em prever a produtividade, indicando os fatores-chave que devem ser priorizados para melhoria ou monitoramento.



6. Predições vs Valores Reais

Este gráfico compara as predições do modelo Random Forest com os valores reais observados no conjunto de teste. Uma boa aproximação das predições aos valores reais indica que o modelo está performando bem na tarefa de previsão da produtividade.



7. Validação Cruzada do Random Forest

A validação cruzada é uma técnica para avaliar a robustez do modelo, dividindo os dados em múltiplos subsets para treinar e testar repetidamente.

O MAE médio obtido na validação cruzada 5-fold do modelo Random Forest foi:

Métrica	Valor
MAE (média ± std)	1.57 ± 0.01