Modelos tradicionais

Hirruá S. da Silva

Introdução

Problema: crescente preocupação com saúde mental no setor de tecnologia; falta de clareza sobre fatores que levam profissionais a buscar tratamento.
Objetivo: construir modelo de classificação supervisionada para identificar preditores da decisão de buscar tratamento.
Dataset: Mental Health in Tech Survey 2014 (Kaggle), com dados demográficos, fatores pessoais (ex: histórico familiar) e fatores do ambiente de trabalho (ex: benefícios, cultura).
Avaliação: classificação binária; métrica inicial: acurácia, além de análise da importância relativa de fatores pessoais vs. organizacionais.

Dados

Análise exploratória: inspeção de tipos de dados, valores nulos, distribuição de variáveis categóricas e balanceamento da variável alvo (treatment).
Outliers: filtragem da coluna Age para manter apenas idades entre 18 e 72 anos.
Inconsistências: normalização de Gender em três categorias: Male, Female e Other.
Dados faltantes: preenchimento com moda em variáveis como self_employed e work_interfere.
Engenharia de atributos: mapeamento binário (yes/no), ordinal encoding (work_interfere, no_employees) e one-hot encoding para variáveis nominais.
Divisão e validação: separação treino (80%) e teste (20%) para evitar data leakage.

Modelos

Algoritmo: RandomForestClassifier.
Avaliação: train/test split (80/20) + Cross-Validation (5 folds).
Baseline: DummyClassifier (estratégia most_frequent) → acurácia 50,5%.

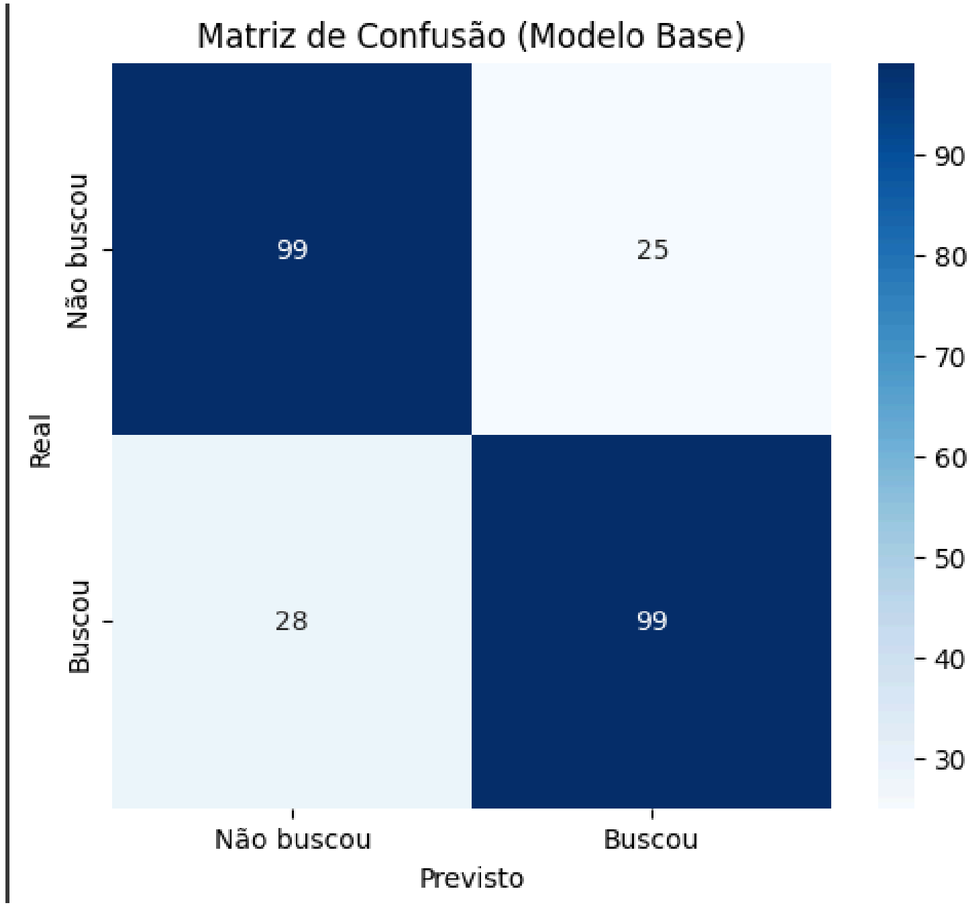
Hiperparâmetros

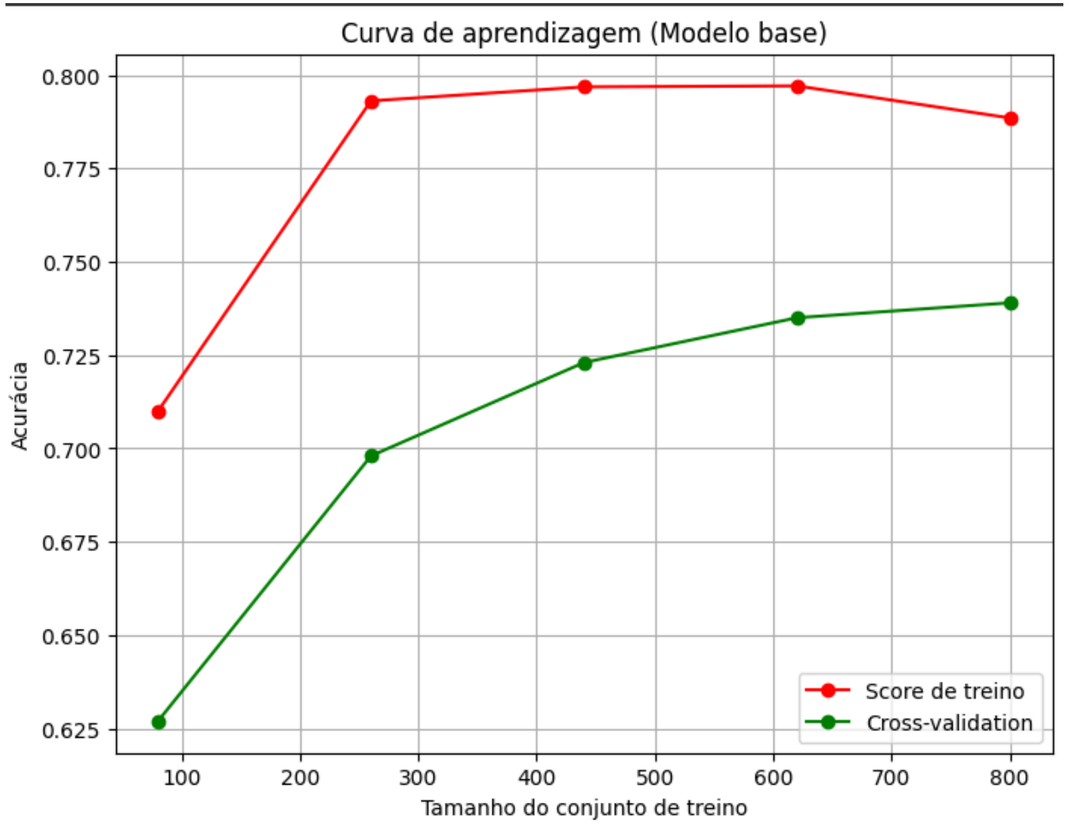
- **n_estimators =** 300
- **max_depth =** 10
- **min_samples_leaf =** 10
- **min_samples_split =** 20
- **max_features =** "sqrt"
- **n_jobs =** -1
- **random_state =** 42

Avaliação

Métricas	RandomForest	DummyClassifier
Acurácia	78,9%	50,6%
F1-SCORE	78,9%	67,2%
ROC-AUC	0.8549	0.5000

Avaliação


Matriz de Confusão (Modelo Base)
Real
Não buscou
Buscou
99
25
28
99
Não buscou
Buscou
Previsto
90
80
70
60
50
40
30

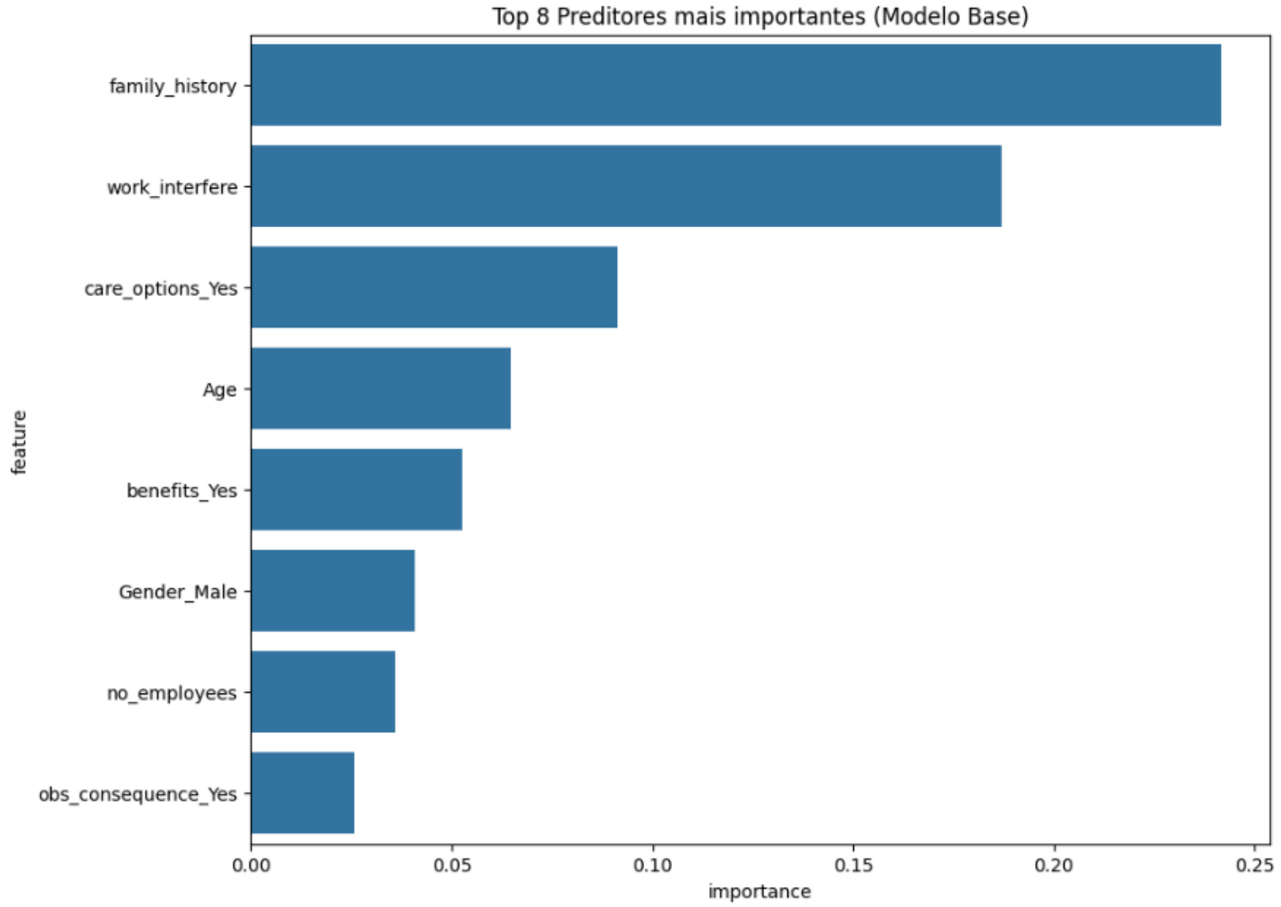

Curva de aprendizagem (Modelo base)
Acurácia
0.800
0.775
0.750
0.725
0.700
0.675
0.650
0.625
100
200
300
400
500
600
700
800
Tamanho do conjunto de treino
Score de treino
Cross-validation

Interpretação

Após a validação da performance e da robustez do modelo RandomForest, a etapa final consiste em interpretar suas decisões para extrair insights e responder à pergunta central do projeto: "**Quais são os principais fatores que influenciam a decisão de um profissional de tecnologia em buscar tratamento para sua saúde mental?**".

Importância de features: calcula a contribuição média de cada variável para a redução do erro de classificação em todas as árvores do ensemble, fornecendo um ranking claro dos preditores mais influentes, como mostra a figura 3.

Interpretação



Refinamento

Refinamento: criação de novas features de interação para capturar relações mais complexas.
Variáveis sintéticas: high_risk_support e support_gap.
Experimento controlado: comparação entre Modelo Base e Modelo Refinado.
Justa comparação: ambos usando RandomForestClassifier com mesmos hiperparâmetros.

Refinamento

Métricas	RandomForest	RandomForest V2
Acurácia	78,9%	76,5%
F1-SCORE	78,9%	76,3%
ROC-AUC	0.8549	0.8526

Conclusão

O projeto buscou construir um modelo para identificar os principais fatores que influenciam a busca por tratamento de saúde mental em profissionais de tecnologia. As tentativas de refinamento, embora metodologicamente válidas, não superaram a performance do modelo inicial, reforçando sua eficácia. A resposta final para a pergunta de pesquisa: "**Quais são os principais fatores que influenciam a decisão de um profissional de tecnologia em buscar tratamento para sua saúde mental?**". Ordem de importância:

- Histórico familiar (family_history)
- A interferência da condição no trabalho (work_interfere)
- Conhecimento sobre opções de cuidado (care_options)