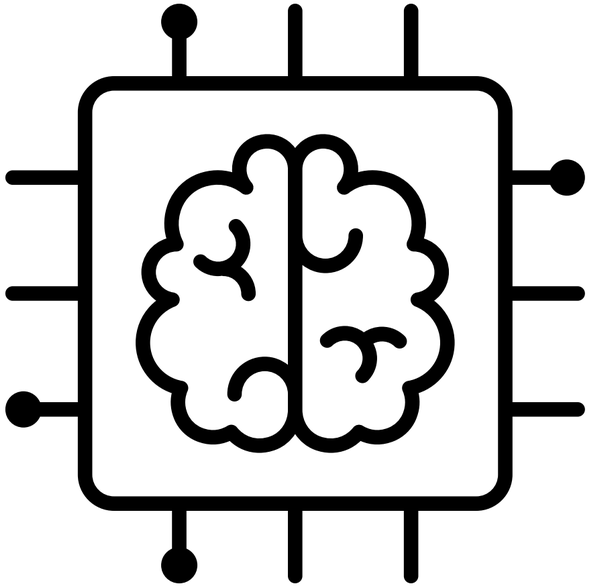
INTELIGÊNCIA ARTIFICIAL - BSI UNESP BAURU

SISTEMA INTELIGENTE

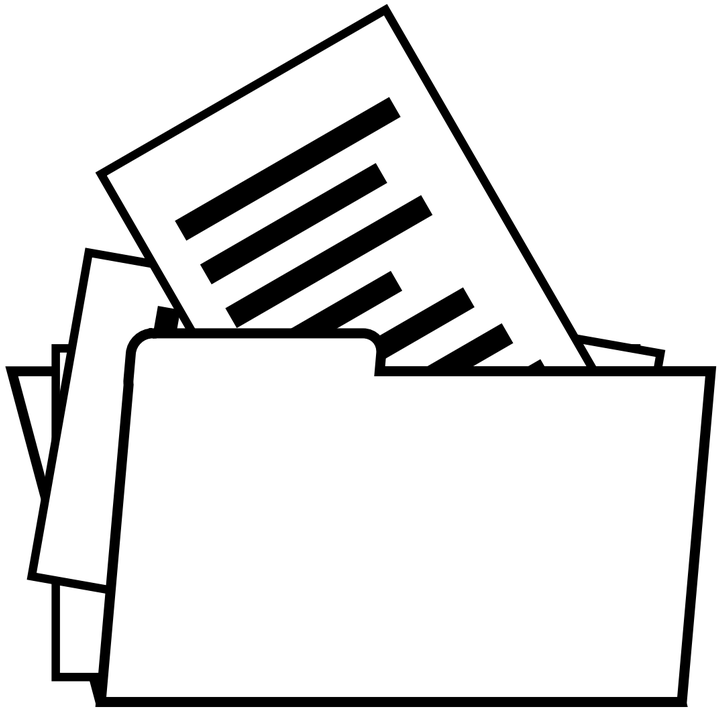
COM DEPLOY EM STREAMLIT



GLÓRIA XAVIER E HELOISA DOS SANTOS

O PROBLEMA

O objetivo principal deste projeto é desenvolver um sistema de previsão que auxilie no diagnóstico de determinadas doenças. O problema central que buscamos resolver é a avaliação
de risco e diagnóstico de saúde, visando auxiliar profissionais da área a entender quais fatores realmente impulsionam o risco de um indivíduo desenvolver uma condição
específica, dada a multiplicidade de diagnósticos possíveis.

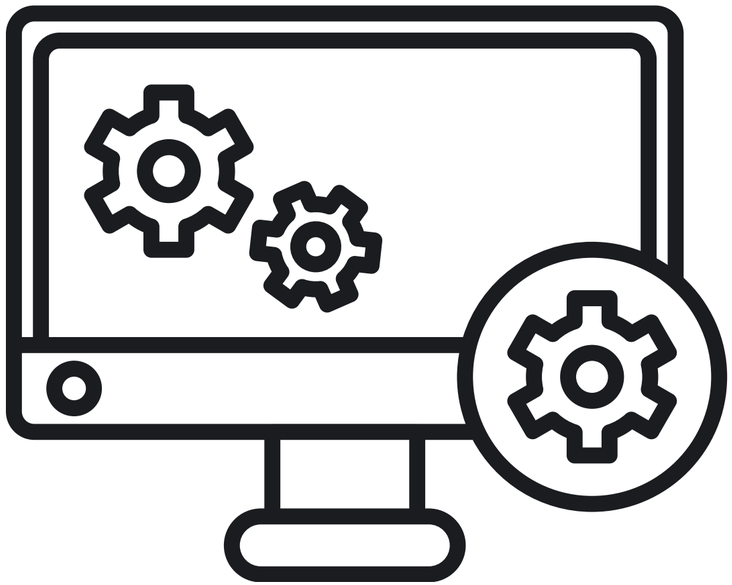


DATASET ESCOLHIDO

Para esta atividade, escolhemos o dataset Healthcare Risk Factors Dataset, disponível no Kaggle.
Este conjunto de dados contém 30.000 registros e 18 variáveis relacionadas às condições de saúde de indivíduos. Ele é adequado para tarefas de classificação em saúde, como prever diabetes, hipertensão e obesidade.

Variável	Tipo de Dado
Age	float64
Gender	object
Medical Condition	object
Glucose	float64
Blood Pressure	float64
BMI	float64
Oxygen Saturation	float64
LengthOfStay	int64
Cholesterol	float64
Triglycerides	float64
HbA1c	float64
Smoking	int64
Alcohol	int64
Physical Activity	float64
Diet Score	float64
Family History	int64
Stress Level	float64
Sleep Hours	float64

MÉTODO UTILIZADO



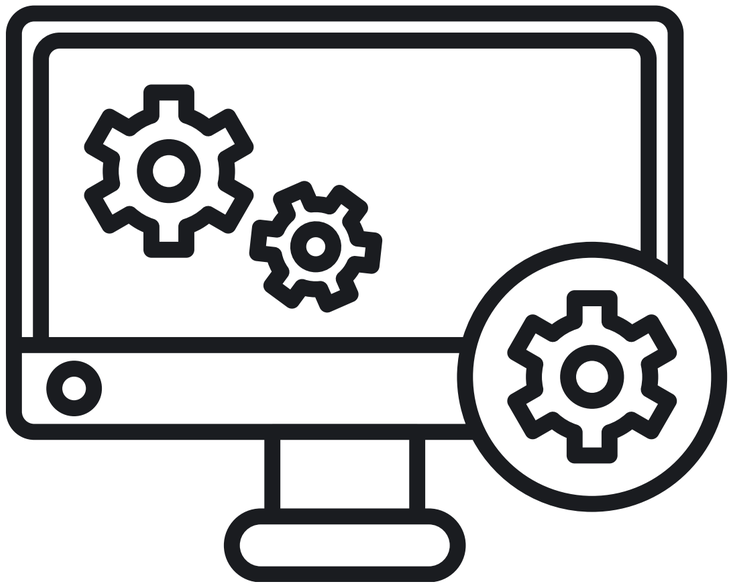
1. **Pré-processamento de Dados:**

- Tratamento de Valores Ausentes: Colunas numéricas foram imputadas com a mediana. Colunas categóricas foram imputadas com a moda. Na coluna da variável alvo, os registros nulos foram deletados
- Codificação de Variáveis Categóricas: A variável alvo (Medical Condition) foi codificada numericamente usando LabelEncoder.

2. **Desenvolvimento e Treinamento do Modelo:**

- Fase 1: Clusterização (K-Means) - O objetivo desta fase foi agrupar os indivíduos em perfis de saúde:
  - 1. Determinação do K: O método do cotovelo (Elbow Method) foi utilizado para identificar o número ideal de clusters, que foi definido como 4.
- 2. Criação do Perfil de Risco: O algoritmo K-Means foi aplicado, e os rótulos de cluster (0 a 3) foram adicionados ao dataset como uma nova variável categórica: Risk_Profile_Cluster.

MÉTODO UTILIZADO



Fase 2: Classificação - O modelo de classificação utilizou as características originais mais o novo Risk_Profile_Cluster para prever a Medical Condition.

- Modelo: Foi utilizado o classificador Random Forest.

1. Divisão de Dados: O conjunto de dados foi dividido em treino e teste (80% treino, 20% teste).
2. Performance do Modelo com K-Means:
    - Acurácia Geral: 90.71%.

Comparação de Performance (Abordagem Híbrida vs. Base)

• Um modelo alternativo, treinado utilizando as features originais sem a variável de cluster, demonstrou uma performance superior, alcançando 91.22% de acurácia.

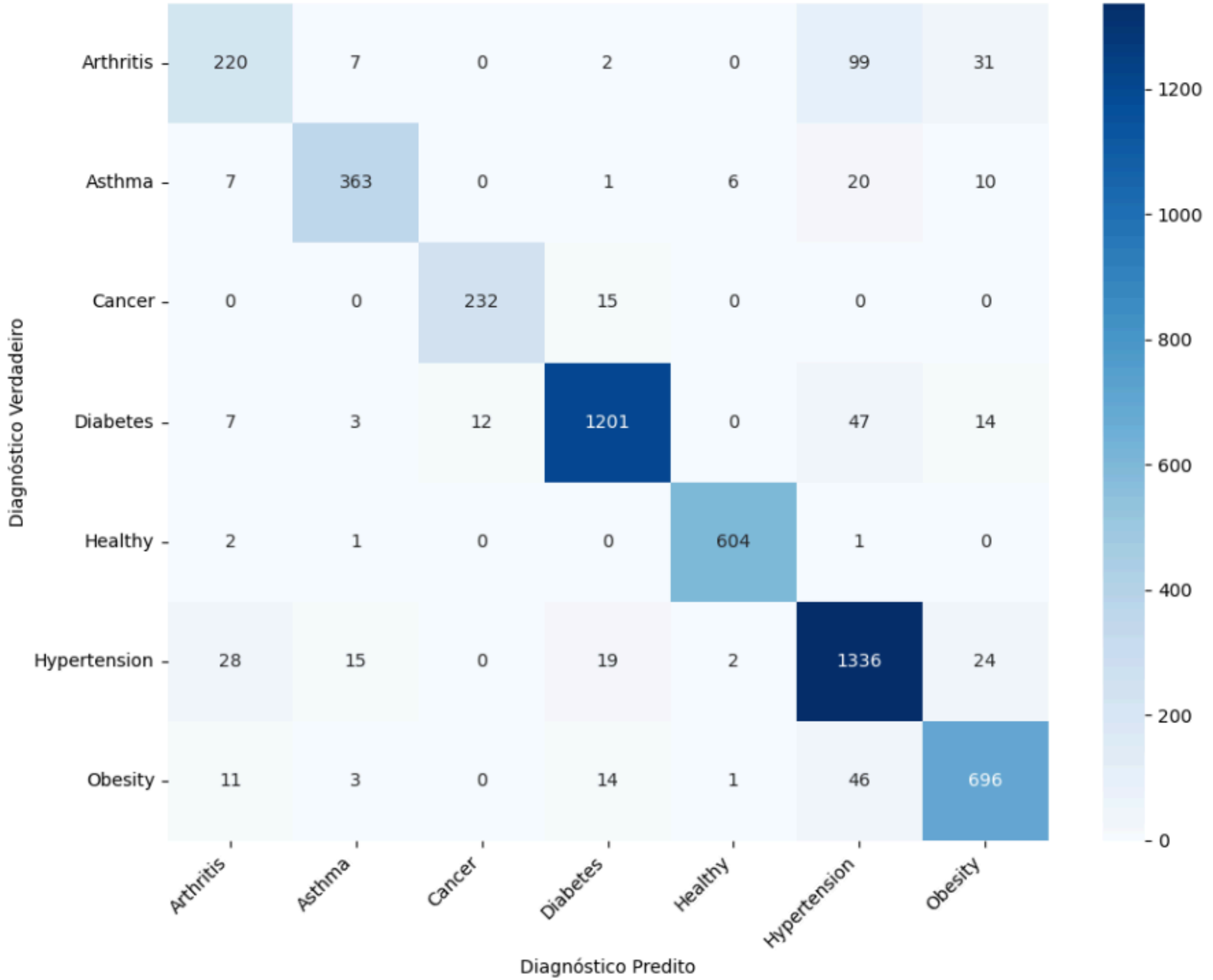
• Devido ao desempenho superior, o Modelo sem Clusterização foi escolhido como o modelo final para o deploy em Streamlit.

MÉTRICAS DE AVALIAÇÃO

Relatório de Classificação

Doença/Métrica	Precision	Recall	F1-Score	Support
Arthritis	0.80	0.61	0.69	359
Asthma	0.93	0.89	0.91	407
Cancer	0.95	0.94	0.95	247
Diabetes	0.96	0.94	0.95	1284
Healthy	0.99	0.99	0.99	608
Hypertension	0.86	0.94	0.90	1424
Obesity	0.90	0.90	0.90	771
---	---	---	---	---
Acurácia Total			0.91	5100
Média Macro	0.91	0.89	0.90	5100
Média Ponderada	0.91	0.91	0.91	5100

Matriz de Confusão



Streamlit

LINK