UNESP – MINERAÇÃO DE DADOS

PREDIÇÃO DE NÍVEL DE DIABETE

Baseada em Hábitos Alimentares e Condições Físicas

Um Estudo com Dados Sintéticos e Reais

VINICIUS S SANTOS

Conteudo

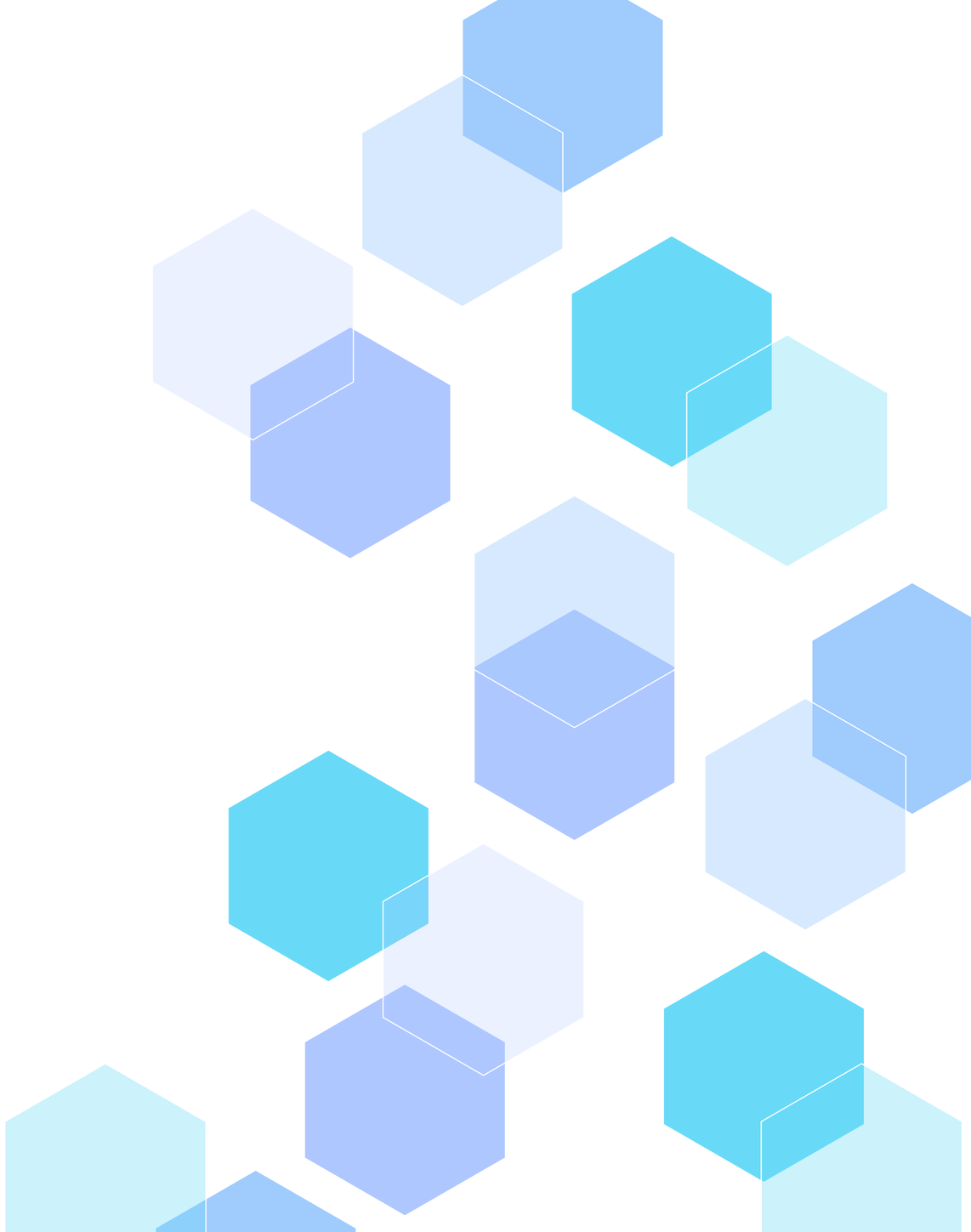
Backlog

Segue abaixo o backlog do Estudo

- **Objetivo do Projeto:**
  - Desenvolver modelos preditivos para classificar níveis de obesidade.
  - Utilizar dados comportamentais e demográficos.
- **Motivação:**
  - Aumento dos índices de obesidade global.
  - Importância do diagnóstico precoce.
- **Tecnologias Utilizadas:**
  - Python, Pandas, Scikit-learn, FastAPI, SMOTE.

Introdução

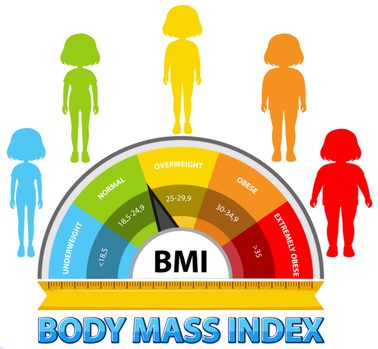
Conceitos sobre objeto de estudo da pesquisa

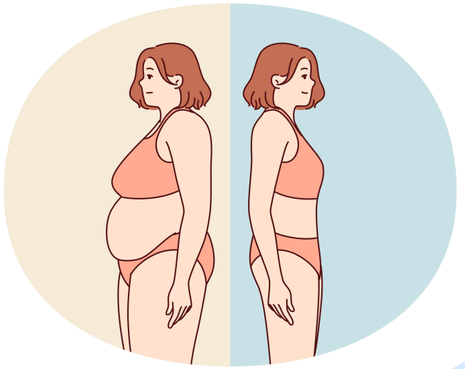


Obesidade

A obesidade é uma doença crônica caracterizada pelo acúmulo excessivo de gordura corporal, resultado de fatores genéticos, ambientais e comportamentais. Ela aumenta o risco de doenças graves, como diabetes tipo 2, hipertensão e problemas cardíacos, destacando a importância do diagnóstico e prevenção.

Atualmente, a obesidade é considerada uma **epidemia global**, afetando milhões de pessoas no mundo todo. A promoção de hábitos saudáveis e a **identificação precoce dos fatores de risco são essenciais para reduzir suas consequências** na saúde pública.





Revisão Sistematica

Em uma Revisão sintetica da Literatura feita pelo IEEE Xplore foi identificado **205 estudos** e apos o tratamento de inclusão e exclusão foi selecionado apaenas **129 estudos relacionados**

Técnica	Estudos
Random Forest	44
Support Vector Machine (SVM)	32
Gradient Boosting	12
Deep Learning	11
Neural Networks	10
Decision Trees	8
XGBoost	8
Bagging	4

Aplicação	Estudos
Predição de Risco de Pacientes	38
Predição de Peso	35
Fatores de Risco	21
Modelos Preditivos	15
Índice de Massa Corporal (IMC)	12
Resultados de Saúde	4
Predição de Risco Geral	2
Predição de Níveis de Obesidade	1
Pacientes Obesos	1

Trabalhos Relacionados

Revista UNIFENAS

Santos, Flávia Aparecida Oliveira, et al. "ESTIMATIVAS DE NÍVEIS DE OBESIDADE UTILIZANDO MACHINE LEARNING: EXPLORANDO FATORES CONTRIBUTIVOS E MODELOS PREDITIVOS PARA A PREVENÇÃO E INTERVENÇÃO NA OBESIDADE." Revista Científica da UNIFENAS-ISSN: 2596-3481 6.5 (2024).

Randon Florest

Dunstan, Jocelyn, et al. "Predicting nationwide obesity from food sales using machine learning." Health informatics journal 26.1 (2020): 652-663.

Arvore de Decisão

DeGregory, Keith W., et al. "A review of machine learning in obesity." Obesity reviews 19.5 (2018): 668-685.

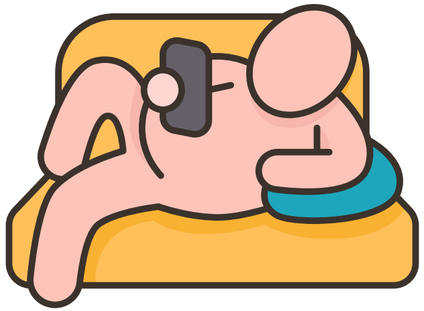
Dataset da Indonesia

Thamrin, Sri Astuti, et al. "Predicting obesity in adults using machine learning techniques: an analysis of Indonesian basic health research 2018." Frontiers in nutrition 8 (2021): 669155.

Dataset estudado

UC Irvine Machine Learning Repository

Mendoza Palechor, Fabio, and Alexis de la Hoz Manotas. "**Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico.**" https://doi. org/10.1016/j. dib. 2019.104344 (2019).



Data description

A base de dados contém 17 variáveis, das quais 16 são características preditivas e 1 é a variável-alvo (target), representando o nível de obesidade do indivíduo.

Demográficas	Medidas Antropométricas	Histórico Familiar e Hábitos Alimentares	Hábitos Comportamentais	Estilo de Vida
Genero	Peso	family_history_with_overweight	Fumante	CALC
Idade	Altura	FAVC e FCVC	CH2O	MTRANS
		NCP	FAF	
		CAEC	TUE	

Pré-Processamento

DataClean

Limpeza de valores nulos/faltantes e valores duplicados

Outliers

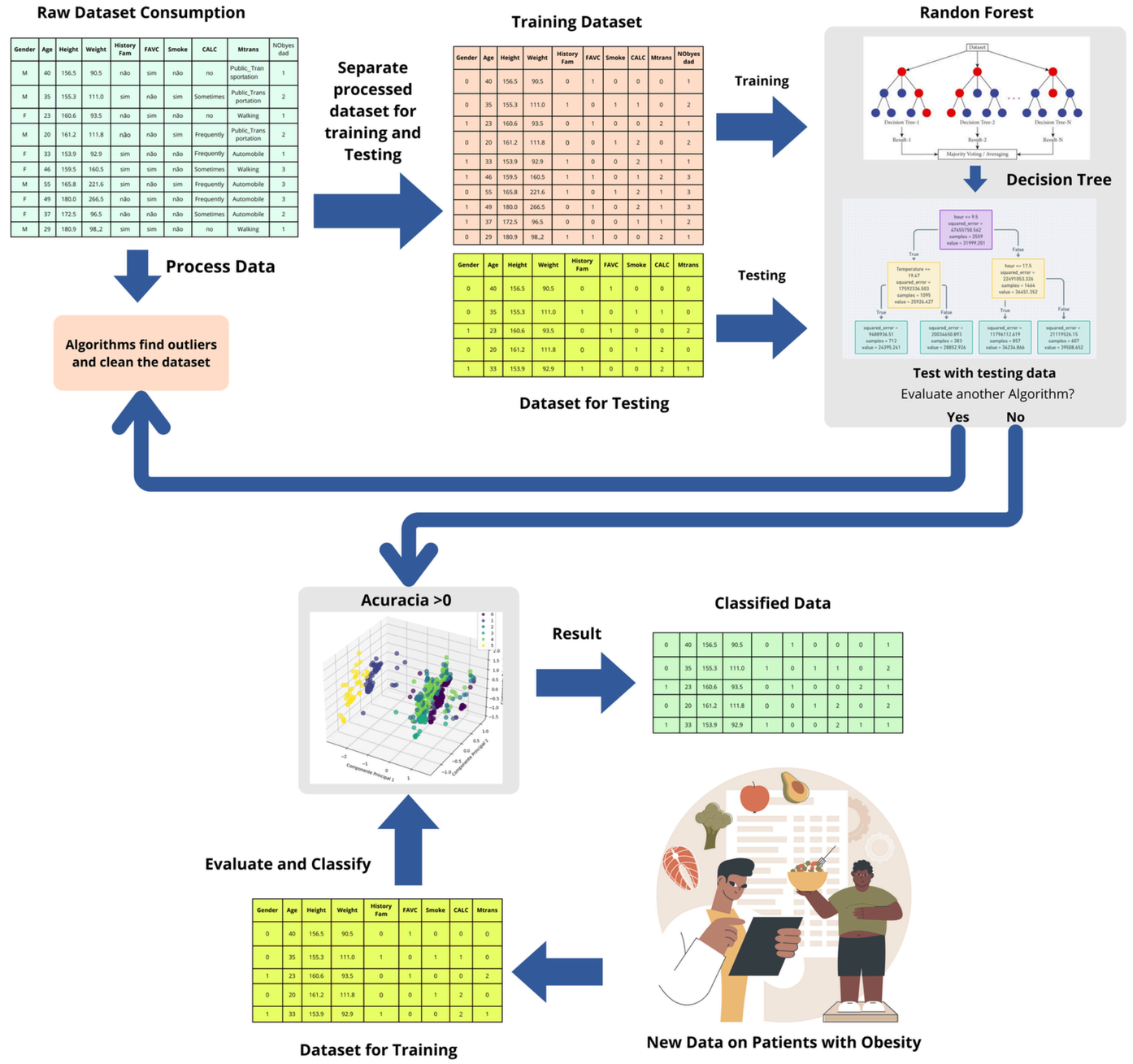
Verificação e limpeza de valores outiliers

Balanceamento das Classes

Balanceamentos das Classes

Metodologia

Foi elaborado o fluxograma a seguir para demonstrar o o desenvolvimento do modelo preditivo do projeto.



Tecnologias Utilizadas

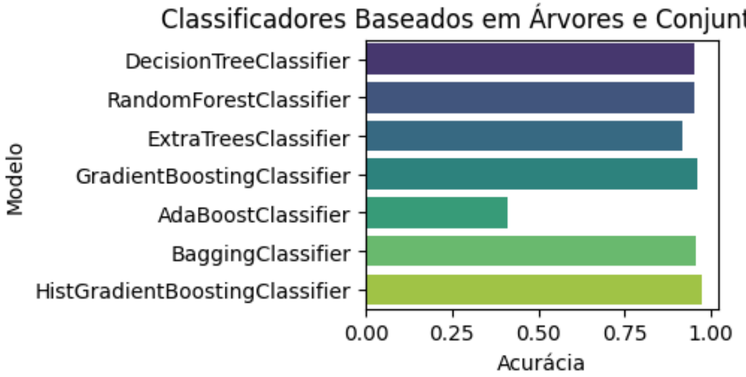
Ambiente	Google Colab
Linguagem de Prog	Python e Javascript
Natureza dos Dados	Sáude
Modelo de Coleta de Dados	Dados Aberto (Online)
Repositorio do Projeto	GITHUB
Periodo	15/09/2024 - 26/11/2024

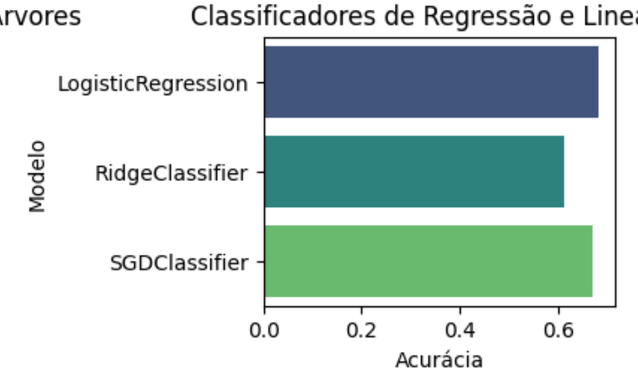
Bibliotecas	Pandas, Sckt-Learn, Seaborn, numpy, fastApi, matplotlib, joblib, pyngrok

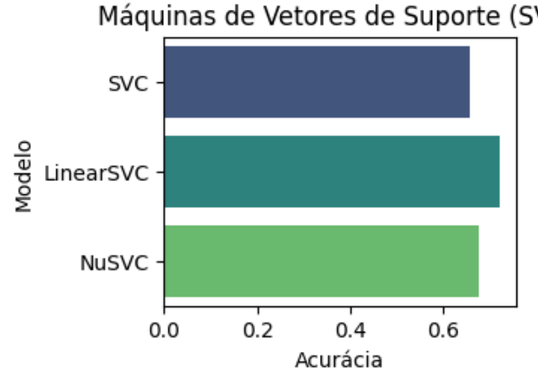
Classificadores do Sckit Learning	StackingClassifier, HistGradientBoostingClassifier, BaggingClassifier, GradientBoostingClassifier, DecisionTreeClassifier, RandomForestClassifier, VotingClassifier, LinearDiscriminantAnalysis, MLPClassifier, KNeighborsClassifier, ExtraTreesClassifier, GaussianProcessClassifier, LinearSVC, LogisticRegression, NuSVC, SVC, SGDClassifier, RidgeClassifier, GaussianNB, CategoricalNB, BernoulliNB, ComplementNB, MultinomialNB, AdaBoostClassifier e QuadraticDiscriminantAnalysis

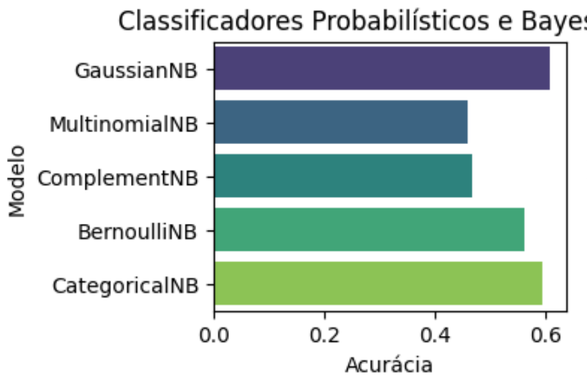
Resultados

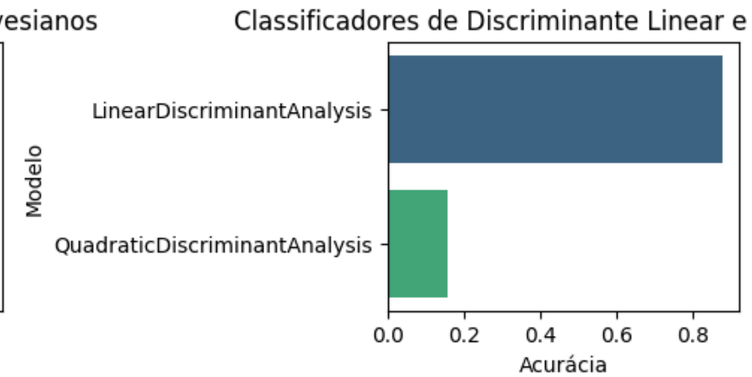
Desempenho Total Desbalanceado

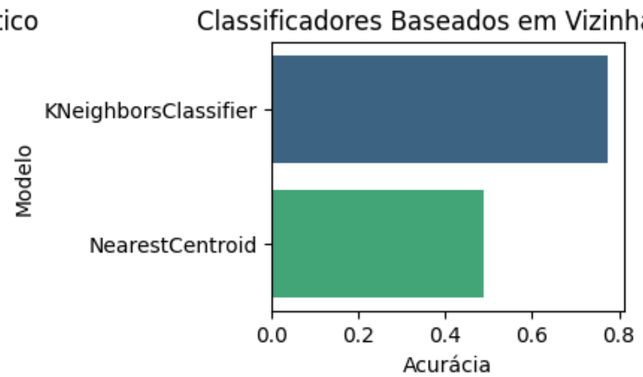


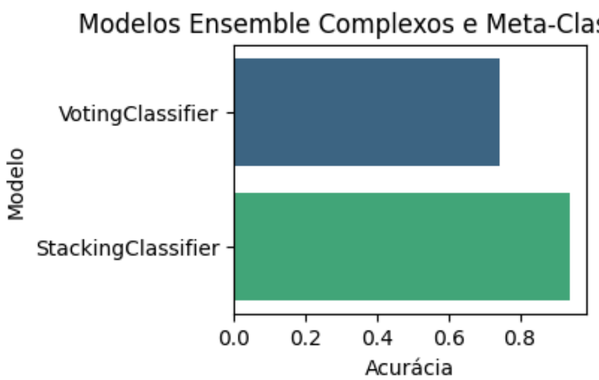


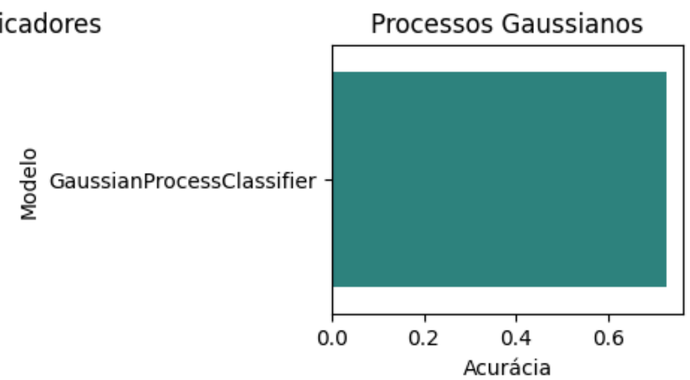


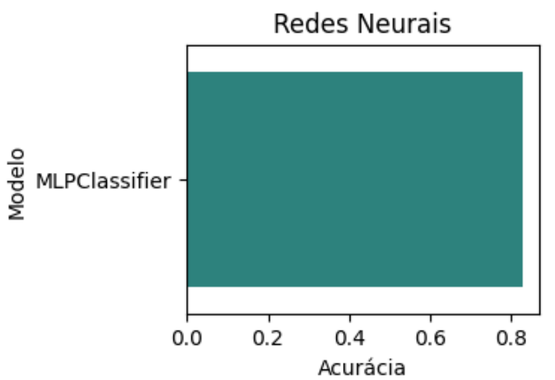




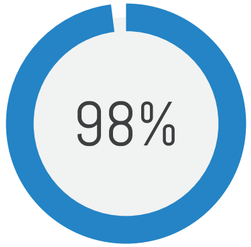






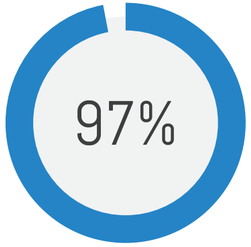


5 Melhores Classificadores



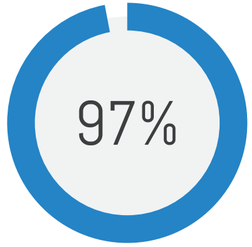
Bagging

Constrói múltiplos modelos (geralmente árvores de decisão) em subconjuntos do conjunto de dados para reduzir a variância e melhorar a estabilidade.



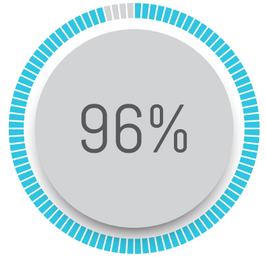
HistGradientBoosting

Variante avançada do Gradient Boosting, otimizada para grandes conjuntos de dados. Combina árvores de decisão para melhorar a precisão



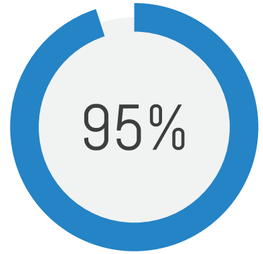
StackingClassifier

Meta-classificador que combina previsões de múltiplos modelos de base, buscando melhorar o desempenho agregando diferentes abordagens.



GradientBoosting

Método de boosting que constrói árvores sequencialmente para corrigir erros de previsões anteriores, aumentando a precisão geral.



RandomForest

Conjunto de árvores de decisão que reduz o overfitting ao combinar previsões de árvores construídas aleatoriamente, sendo robusto e eficiente em diversos cenários.

Comparação dos 5 Classificadores

HistGradientBoosting e StackingClassifier

O HistGradientBoosting apresentou o melhor desempenho geral, com alta acurácia, F1 Score, precisão e recall, adequado para capturar padrões complexos nos dados. Já o StackingClassifier combina a força de múltiplos classificadores, trazendo versatilidade e boa performance, mas com maior custo computacional.

GradientBoosting, Bagging e RandomForest

O GradientBoosting oferece precisão ajustando erros iterativamente, enquanto o BaggingClassifier utiliza amostragem para aumentar a estabilidade, ideal para dados ruidosos. O RandomForest , composto por múltiplas árvores de decisão, proporciona excelente generalização e é eficiente para grandes conjuntos de dados.