Aprendizado Estatístico em Dados Longitudinais

Cícero Hitzschky

cicero.hitzschky@alu.ufc.br

Departamento de Estatística e Matemática Aplicada
Universidade Federal do Ceará

27 de fevereiro de 2025

UNIVERSIDADE FEDERAL DO CEARÁ

Sumário

UNIVERSIDADE FEDERAL DO CEARÁ

O que é Inteligência Artificial (IA)?

- Não há uma única forma de definir;
- Diferentes autores a descrevem de maneiras distintas a depender do contexto em que está sendo empregada.

O que é IA?

Figura: Prof. John Haugeland

- *Artificial Intelligence: The Very Idea (1985).*
- "O novo e interessante esforço para fazer os computadores pensarem (...) máquinas com mentes, no sentido total e literal."

O que é IA?

- *Artificial Intelligence (1972).*
- "[Automatização de] atividades que associamos ao pensamento humano, atividades como a tomada de decisões, a resolução de problemas, o aprendizado..."

Figura: Richard Bellman

O que é IA?

Figura: Raymond Kurzweil

- *The Age of Intelligent Machines (1990).*
- "A arte de criar máquinas que executam funções que exigem inteligência quando executadas por pessoas.""

O que é IA?

- “ O estudo das faculdades mentais pelo uso de modelos computacionais. ” (Charniak e McDermott, 1985)
- “ O estudo das computações que tornam possível perceber, raciocinar e agir. ” (Winston, 1992)
- “ O estudo de como os computadores podem fazer tarefas que hoje são melhor desempenhadas pelas pessoas. ” (Rich and Knight, 1991)
- “ “Inteligência Computacional é o estudo do projeto de agentes inteligentes. ” (Poole et al., 1998)

O que é IA?

Segundo a Data Science Academy (DSA)

"Inteligência Artificial (IA) refere-se ao campo da ciência da computação que se concentra na criação de sistemas capazes de realizar tarefas que normalmente exigiriam inteligência humana."

Inteligência Artificial VS Aprendizado Estatístico

Definição

"Aprendizado de Máquina [ou Estatístico] é o campo de estudo da ciência da computação que dá ao computador a habilidade de aprender sem ser explicitamente programado." Arthur Samuel (1959)

O que é aprender?

Se baixássemos todo o conteúdo da Wikipédia em um computador, poderíamos afirmar que ele aprendeu?

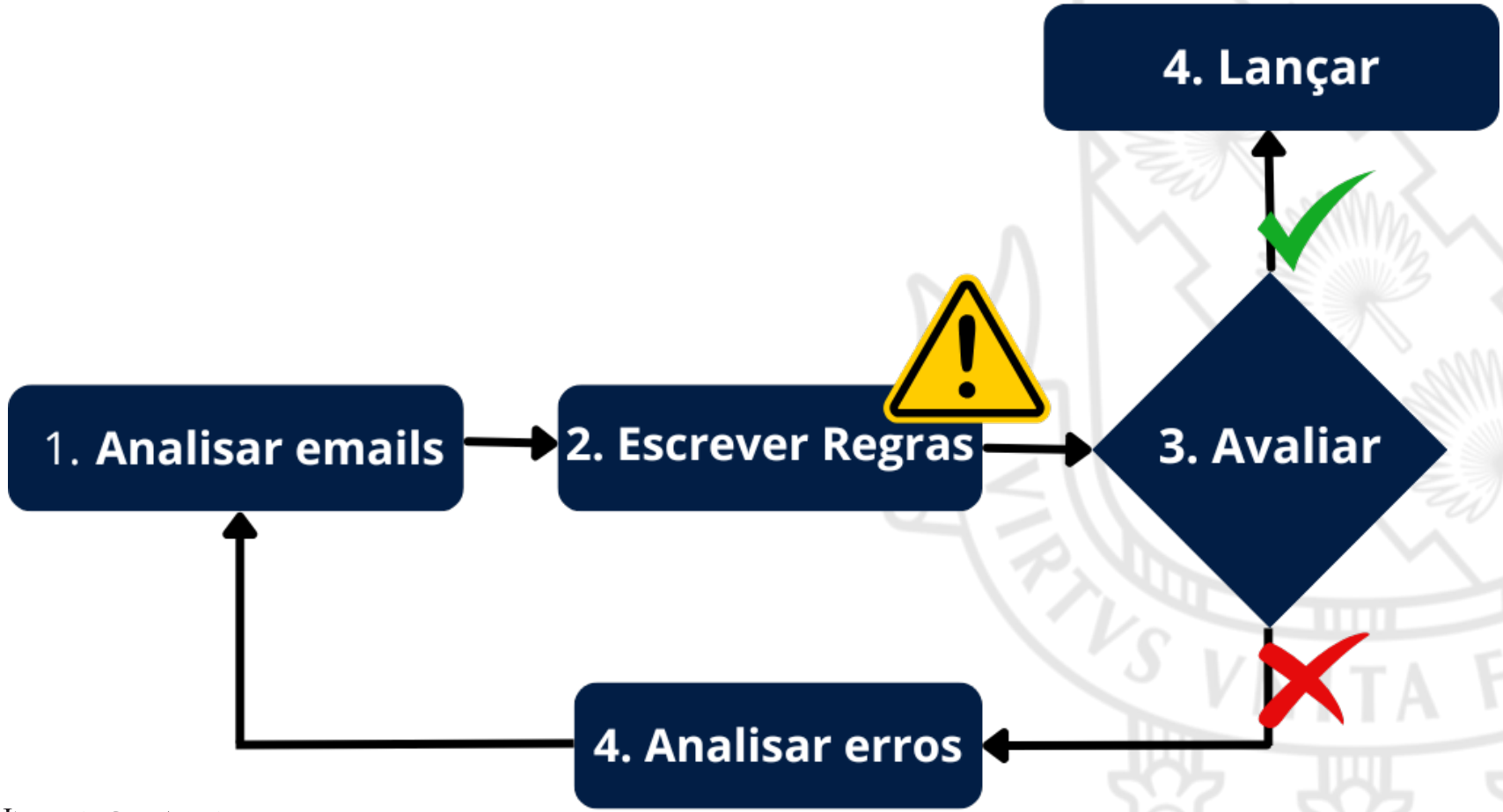
Inteligência Artificial VS Aprendizado Estatístico

Exemplo

Suponha que desejamos criar um filtro de *spam*. Uma forma de fazer isso seria seguindos os passos:

1. Examinar emails que sejam *spam*. É provável que repitam-se as frases: *para você*, *gratuito*, *oportunidade*, *imperdível*...
2. Escrever um programa que busque por uma das palavras citadas acima no corpo do email e acione uma *flag*.
3. Testamos o programa e repetimos os passos 1 e 2 até nosso programa estar bom o suficiente.

Inteligência Artificial VS Aprendizado Estatístico



Inteligência Artificial VS Aprendizado Estatístico

Observações

- Essa solução funciona? **SIM!!** 😄;
- É a melhor forma de fazer? **NÃO!!** 😑;
- Sensível a pequenas mudanças, e.g, Pra Você, PARA VOCÊ, PRA VOCE...

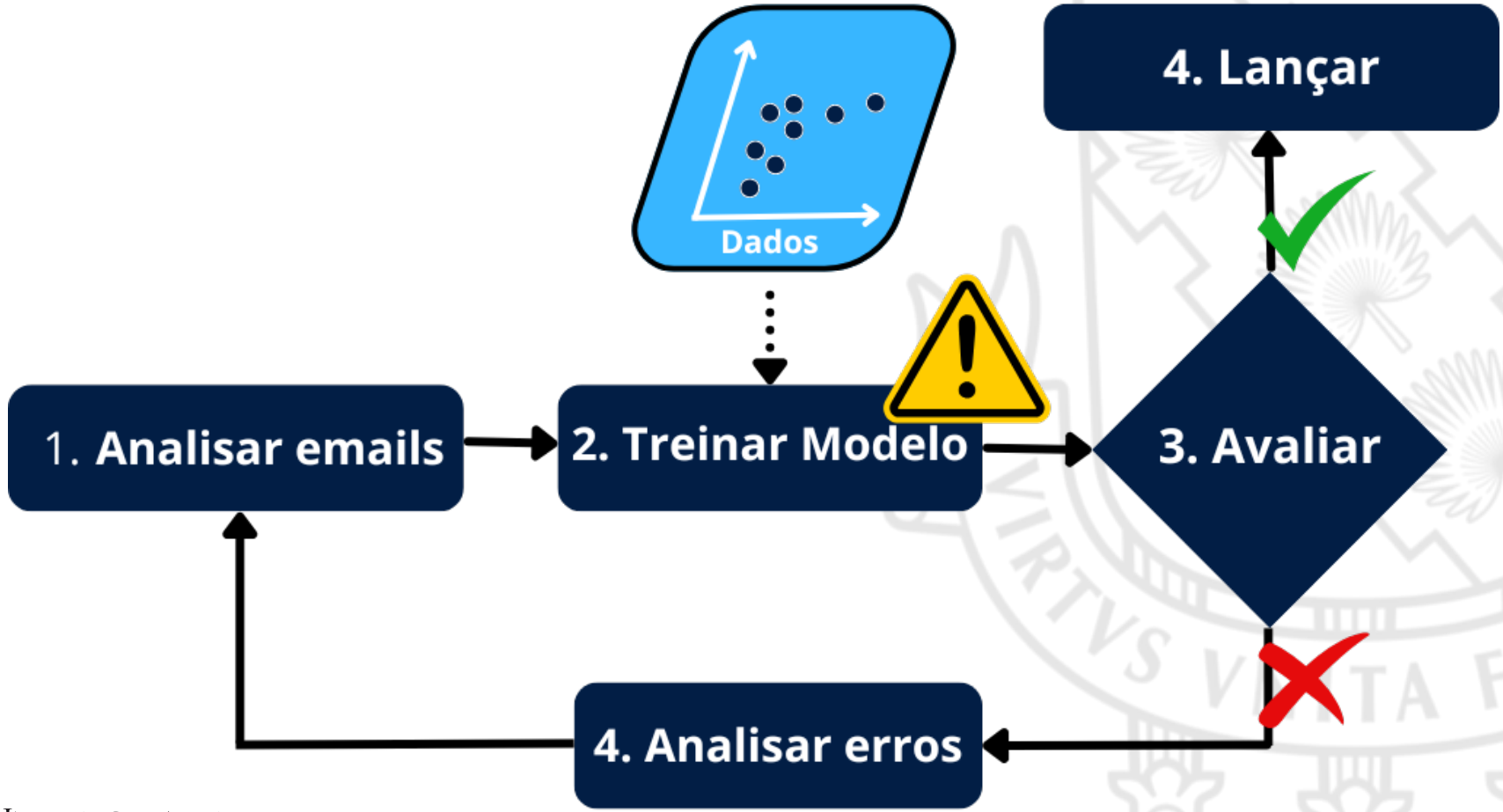
Inteligência Artificial VS Aprendizado Estatístico

Exemplo *(Filtro de Spam)*

Outra abordagem seria:

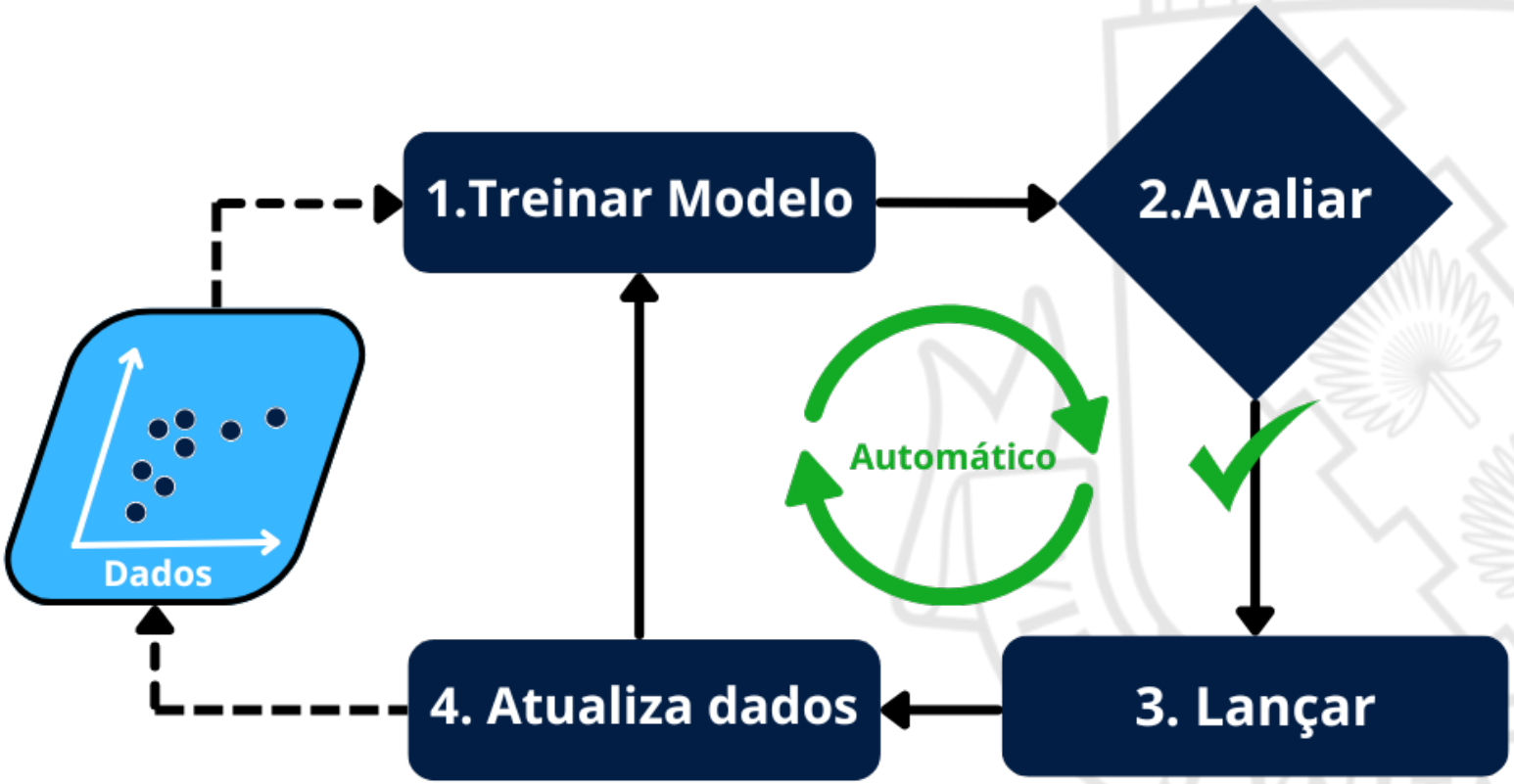
1. Examinar emails que sejam *spam*;
2. Treinar um modelo de aprendizado de máquina para classificar emails como *spam* ou *ham*.

Inteligência Artificial VS Aprendizado Estatístico



UNIVERSIDADE FEDERAL DO CEARÁ

Inteligência Artificial VS Aprendizado Estatístico



UNIVERSIDADE FEDERAL DO CEARÁ

Possíveis complicações com abordagem Estatística Tradicional em Dados Longitudinais

- Geralmente exigem técnicas mais avançadas;
- Herança de suposições, e.g, homoscedasticidade, normalidade, independência de observações, falta correlação entre variáveis preditoras...

Suposições de alguns modelos estatísticos

Modelos de Efeitos Fixos

- Independência Condicional: Dados os efeitos fixos, os erros são independentes;
- Ausência de Correlação Serial;
- O erro aleatório é não correlacionado ao longo do tempo;
- Homoscedasticidade dos resíduos.

Suposições de alguns modelos estatísticos

Modelos de Efeitos Aleatórios

- Normalidade dos componentes aleatórios bem como dos resíduos;
- Homoscedasticidade;
- Ausência de correlação serial.

Suposições de alguns modelos de Aprendizado Estatísticos

Naive Bayes

Independência dos preditores.

Floresta Aleatória

Nenhum dado ausente.

Redes Neurais

Nenhum dado ausente.

Gradient Boosting

Nenhuma suposição inicial.

UNIVERSIDADE FEDERAL DO CEARÁ

Comparação: Estatística Tradicional vs. Aprendizado Estatístico

Atributo	**Estatística Tradicional**	**Aprendizado de Máquina**
Adequado para	Teste de hipóteses dedutivas	Geração de hipóteses abdutivas
Forma das relações entre variáveis	Ajusta os dados em formas pre-definidas especificadas no modelo estatístico	Aprende a verdadeira forma da relação entre variáveis
Objetivo principal	Testar se existem relacionamentos pré-especificados nos dados	Identificar padrões nos dados sem preconceitos
Padrão geral de resultados	Baixa previsibilidade, alta explica-bilidade	Alta previsibilidade, baixa explica-bilidade

Comparação: Estatística Tradicional vs. Aprendizado Estatístico

Atributo	**Estatística Tradicional**	**Aprendizado de Máquina**
Como confiar na análise	Testes de robustez	Modelo de teste em dados não vistos. Teste a generalização por meio de análise secundária
Experiência necessária	Experiência em modelos estatísticos	Experiência em análise de diversos conjuntos de dados
Habilidades de pesquisador	Treinamento em estatística	Treinamento em ciência de dados e programação

Experiência e Poder Computacional

Atributo	**Estatística Tradicional**	**Aprendizado de Máquina**
Poder computacional Exigido	Laptops modernos geralmente são suficientes	Requer ambiente de computação de ponta
Reutilização de modelos	Precisa construir modelos diferentes para cada objetivo	Um algoritmo pode ser reutilizado para diferentes objetivos
Número de preditores	Limitado pela correlação entre variáveis	Limitado pelo poder computacional

Contextualização

Osteoporose

é uma doença em que a degradação estrutural e a diminuição da densidade mineral dos ossos (DMO) aumentam o risco de fraturas ósseas.

Motivações para o estudo

- A prevalência da osteoporose aumentou drasticamente nos últimos anos;
- Representa um problema de saúde pública com alto índice de morbidade (muito comum).

Contextualização

- Em 2010, na União Europeia, aproximadamente 5,5 milhões de homens e 22 milhões de mulheres foram afetados pela osteoporose;
- 80% das mulheres afetadas não estavam cientes dos seus fatores de risco até o diagnóstico.

Objetivos

1. Comparar a precisão preditiva dos métodos aprendizado de máquina com a regressão linear múltipla tradicional;
2. Classificar a importância de vários fatores de risco, incluindo dados demográficos, estilo de vida e bioquímica, na previsão das mudanças futuras no δ-T score.

Metodologia

UNIVERSIDADE FEDERAL DO CEARÁ

Fonte dos Dados

- Coorte MJ de Taiwan
  - Coorte prospectiva em andamento
  - Exames conduzidos pelos Centros de Triagem de Saúde MJ

Informações Coletadas

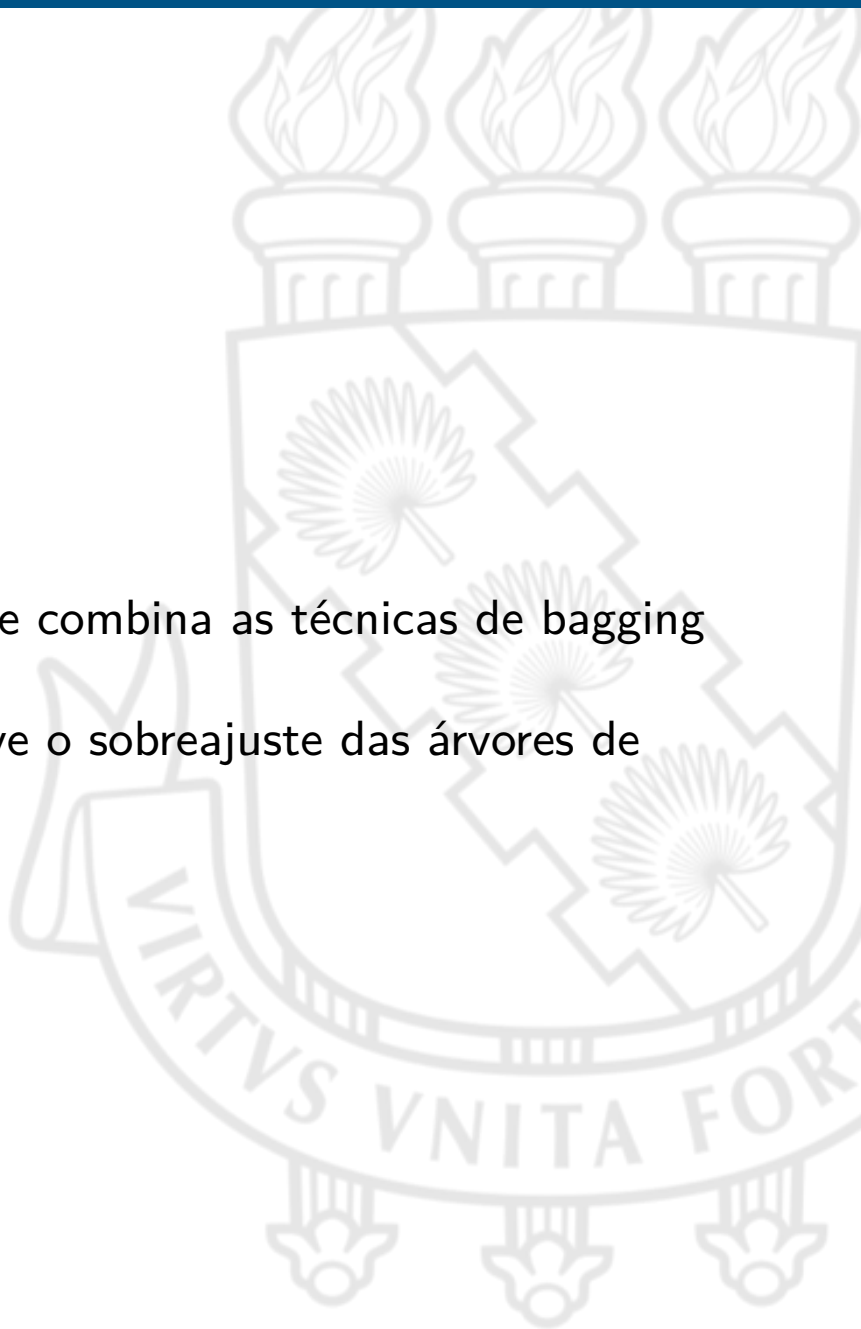
- Mais de 100 indicadores biológicos essenciais
- Questionário abrangendo:
  - Histórico médico pessoal e familiar
  - Estado de saúde atual
  - Estilo de vida e exercício físico
  - Hábitos de sono e alimentares

Considerações Éticas

- Consentimento informado dos participantes
- Aprovação pelo Comitê de Ética

Modelos Utilizados

- Floresta Aleatória (RF)
  - É baseado em árvores de decisão que combina as técnicas de bagging e boosting.
  - Minimiza a função de perda e resolve o sobreajuste das árvores de decisão tradicionais.



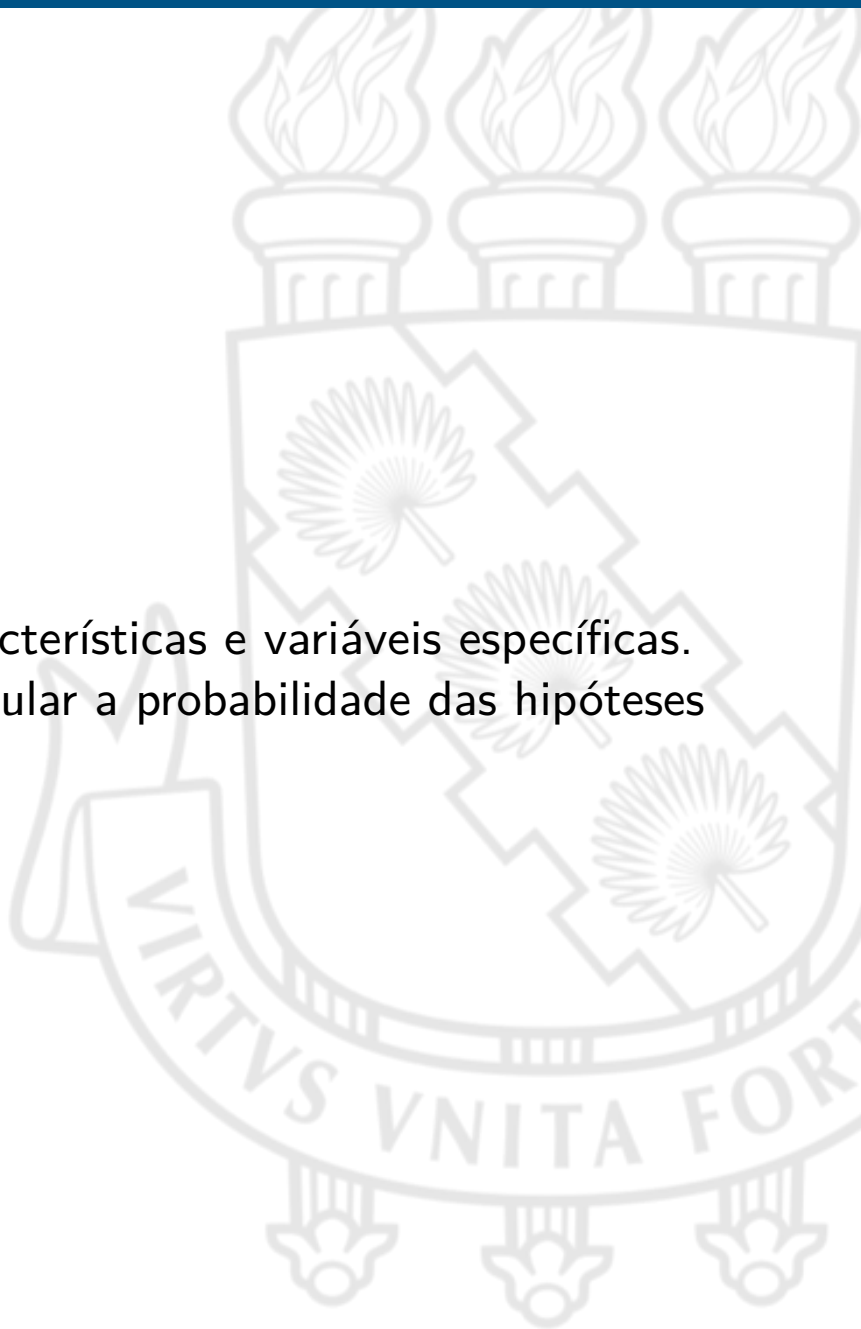
Modelos Utilizados

- Gradient Boosting Estocástico (SGB)
  - Classifica objetos com base em características e variáveis específicas.
  - Utiliza o teorema de Bayes para calcular a probabilidade das hipóteses sobre grupos presumidos.

UNIVERSIDADE FEDERAL DO CEARÁ

Modelos Utilizados

- Naive Bayes (NB)
  - Classifica objetos com base em características e variáveis específicas.
  - Utiliza o teorema de Bayes para calcular a probabilidade das hipóteses sobre grupos presumidos.



Modelos Utilizados

- Extreme Gradient Boosting (XGBoost)
  - Tecnologia de gradient boosting baseada na extensão otimizada do SGB.
  - Treina vários modelos “fracos” e faz ensemble com o Gradiente Boosting.

Métricas de Desempenho

Métrica	Equação
Erro percentual médio simétrico absoluto	$\text{SMAPE} = \frac{1}{n}\sum_{t=1}^{n}\frac{\lvert y_t - \hat{y}_t\rvert}{(\lvert y_t\rvert + \lvert\hat{y}_t\rvert)/2} \times 100$
Erro absoluto relativo	$\text{RAIE} = \frac{\sum_{t=1}^{n}\lvert\hat{y}_t - y_t\rvert}{\sum_{t=1}^{n}\lvert y_t - \bar{y}\rvert}$
Raiz do Erro Quadrático Relativo	$\text{RRSE} = \sqrt{\frac{\sum_{t=1}^{n}(\hat{y}_t - y_t)^2}{\sum_{t=1}^{n}(y_t - \bar{y})^2}}$
Raiz do Erro Quadrático Médio	$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(\hat{y}_t - y_t)^2}$

Comparação dos métodos

Modelo	SMAPE	RAE	RRSE	RMSE
LR	1,3601	1,0617	1,0459	0,6107
RF	1,3246	**1,0185**	1,0242	0,5980
SGB	**1,2764**	1,0193	**1,0209**	**0,5961**
XGBoost	1,3601	1,0616	1,0459	0,6107
NB	1,3385	1,0330	1,0281	0,6003

Resultados do Estudo

- O estudo demonstrou que todos os quatro métodos Mach-L superaram a regressão linear múltipla (MLR) tradicional.
- Desempenho dos métodos Mach-L:
  - DBP (Pressão Diastólica)
  - SBP (Pressão Sistólica)
  - UA (Ácido Úrico)
  - Nível de escolaridade
  - TG (Triglicerídeos)
  - Horas de sono
- Identificação dos fatores de risco mais importantes.

Referências I

Obrigado!!!

Contato:

cicero.hitzschky@alu.ufc.br

UNIVERSIDADE FEDERAL DO CEARÁ