# Aplicando Árvore de Decisão e Naive Bayes na Previsão da Fertilidade

Victor Cardel, Luis Modesto

Universidade Federal da Bahia

3/12/2018

# Introdução

- **Proposta:** Construir e avaliar um modelo de machine learning que tenta predizer se um determinado casal, dadas certas características, tem ou não boas chances de conceberem um filho.

# Introdução

- **Proposta:** Construir e avaliar um modelo de machine learning que tenta predizer se um determinado casal, dadas certas características, tem ou não boas chances de conceberem um filho.
- Pode ser pensado como um problema de **classificação**.

# Introdução

- **Proposta:** Construir e avaliar um modelo de machine learning que tenta predizer se um determinado casal, dadas certas características, tem ou não boas chances de conceberem um filho.
- Pode ser pensado como um problema de **classificação**.
- Foi utilizado o paradigma **supervisionado** utilizando os algoritmos **Naive Bayes** e **Árvore de Decisão**.

# Dados

- **PALS** (Pregnancy and Lifestyle Study).

# Dados

- **PALS** (Pregnancy and Lifestyle Study).
- Dataset aberto fornecido pela University of South Australia.

# Dados

- **PALS** (Pregnancy and Lifestyle Study).
- Dataset aberto fornecido pela University of South Australia.
- Contem uma grande variedade de parâmetros em relação à saúde, fertilidade e hábitos de casais que planejavam ter um filho.

# Dados

- **PALS** (Pregnancy and Lifestyle Study).
- Dataset aberto fornecido pela University of South Australia.
- Contem uma grande variedade de parâmetros em relação à saúde, fertilidade e hábitos de casais que planejavam ter um filho.
- Tabela original com **565 instâncias**, cada uma contendo **494 atributos**.

# Dados

- **PALS** (Pregnancy and Lifestyle Study).
- Dataset aberto fornecido pela University of South Australia.
- Contem uma grande variedade de parâmetros em relação à saúde, fertilidade e hábitos de casais que planejavam ter um filho.
- Tabela original com **565 instâncias**, cada uma contendo **494 atributos**.
- Na tabela havia a informação se o casal tinha conseguido engravidar ou não após seis meses de tentativa.

# Dados

| A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|
| Couple ID | dateintm | MaleAge | questcm | cobm | ymigm | educasm | educatm | occup1m |
| P0001f | 6/30/1988 | 23 | 1 | 1 | 0 | 1 | 2 | 2207 |
| P0004f | 7/8/1988 | 32 | 1 | 2 | 66 | 1 | 2 | 2213 |
| P0005f | 7/11/1988 | 27 | 1 | 2 | 62 | 1 | 12 | 2 |
| P0006f | 7/15/1988 | 30 | 1 | 1 | 0 | 2 | 2 | 5501 |
| P0008f | 7/20/1988 | 34 | 3 | 1 | 0 | 1 | 0 | 5305 |
| P0009f | 8/1/1988 | 31 | 2 | 1 | 0 | 1 | 2 | 5301 |
| P0010f | 8/8/1988 | 29 | 2 | 1 | 0 | 2 | 1 | 1201 |
| P0011f | 8/9/1988 | 26 | 1 | 1 | 0 | 1 | 0 | 5601 |
| P0012f | 8/8/1988 | 30 | 2 | 1 | 0 | 1 | 2 | 2403 |
| P0014f | 4/3/1989 | 36 | 3 | 1 | 0 | 1 | 2 | 2601 |
| P0015f | 9/15/1988 | 42 | 2 | 1 | 0 | 1 | 2 | 4 |
| P0016f | 9/16/1988 | 25 | 2 | 2 | 79 | 2 | 1 | 4705 |
| P0018f | 4/10/1988 | 28 | 2 | 1 | 0 | 1 | 0 | 7105 |
| P0021f | 9/22/1988 | 32 | 2 | 1 | 0 | 3 | 0 | 8499 |
| P0022f | 9/28/1988 | 34 | 2 | 2 | 84 | 1 | 2 | 2807 |
| P0024f | 10/3/1988 | 36 | 2 | 1 | 0 | 1 | 2 | 2703 |
| P0025f | 9/27/1988 | 30 | 2 | 1 | 0 | 2 | 1 | 4311 |
| P0028f | 11/4/1988 | 27 | 3 | 1 | 0 | 1 | 0 | 5503 |
| P0029f | 10/24/1988 | 43 | 3 | 5 | 58 | 3 | 0 | 8921 |
| P0030f | 10/18/1988 | 31 | 2 | 1 | 0 | 1 | 1 | 1601 |
| P0031f | 10/7/1988 | 27 | 2 | 1 | 0 | 2 | 1 | 4101 |
| P0033f | 9/26/1988 | 28 | 3 | 7 | 78 | 1 | 2 | 2 |

# Pré-Processamento

- Número de atributos considerados foram reduzidos.

# Pré-Processamento

- Número de atributos considerados foram reduzidos.
- Tanto para o homem quanto para a mulher foram selecionados:

# Pŕe-Processamento

- Idade.

# Pŕe-Processamento

- Idade.
- Se praticava exercícios ou não.

# Pŕe-Processamento

- Idade.
- Se praticava exercícios ou não.
- Se fumava ou não.

# Pŕe-Processamento

- Idade.
- Se praticava exercícios ou não.
- Se fumava ou não.
- Se bebia ou não.

# Pŕe-Processamento

- Idade.
- Se praticava exercícios ou não.
- Se fumava ou não.
- Se bebia ou não.
- Se usava alguma substância ilícita ou não.

# Pŕe-Processamento

- Idade.
- Se praticava exercícios ou não.
- Se fumava ou não.
- Se bebia ou não.
- Se usava alguma substância ilícita ou não.
- Se tinha problema reprodutivo ou não.
- Se fez vasectomia (ou ligação das tubas) ou não.

# Pŕe-Processamento

- Idade.
- Se praticava exercícios ou não.
- Se fumava ou não.
- Se bebia ou não.
- Se usava alguma substância ilícita ou não.
- Se tinha problema reprodutivo ou não.
- Se fez vasectomia (ou ligação das tubas) ou não.
- Se foi diagnosticado com alguma DST ou não.

# Pŕe-Processamento

- Idade.
- Se praticava exercícios ou não.
- Se fumava ou não.
- Se bebia ou não.
- Se usava alguma substância ilícita ou não.
- Se tinha problema reprodutivo ou não.
- Se fez vasectomia (ou ligação das tubas) ou não.
- Se foi diagnosticado com alguma DST ou não.
- Rótulo (conseguiram ter um filho ou não).

# Pŕe-Processamento

- Idade.
- Se praticava exercícios ou não.
- Se fumava ou não.
- Se bebia ou não.
- Se usava alguma substância ilícita ou não.
- Se tinha problema reprodutivo ou não.
- Se fez vasectomia (ou ligação das tubas) ou não.
- Se foi diagnosticado com alguma DST ou não.
- Rótulo (conseguiram ter um filho ou não).
- Todos os atributos numéricos foram transformados em categóricos e dados faltantes foram preenchidos com a moda.

# Tabela Processada

| A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | Q | P | R | engravidou |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| IF | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | IF | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| IF | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | IF | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| IF | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | IF | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| IF | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | IF | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| IF | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | IF | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| IF | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | IF | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| IF | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | IF | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| IF | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | IF | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| IF | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | IF | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 |
| IF | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | II | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 |
| IF | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | II | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| IF | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | II | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| IF | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | IF | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| IF | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | II | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 0 |
| IF | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | IF | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 |
| IF | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | IF | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 |
| IF | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | IF | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 |
| IF | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | IF | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0 |
| IF | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | II | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0 |
| IF | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | IF | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| IF | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | IF | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| IF | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | IF | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| IF | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | IF | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 0 |
| IF | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | IF | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| IF | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | IF | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 |
| IF | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | IF | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 |

# Tabela Processada

Tabela original reduzida para uma tabela com 18 atributos categóricos e 565 instâncias.

# Naive Bayes

# Árvore de Decisão

# Validação

- Foi escolhido o método de validação **10-fold cross validation**.

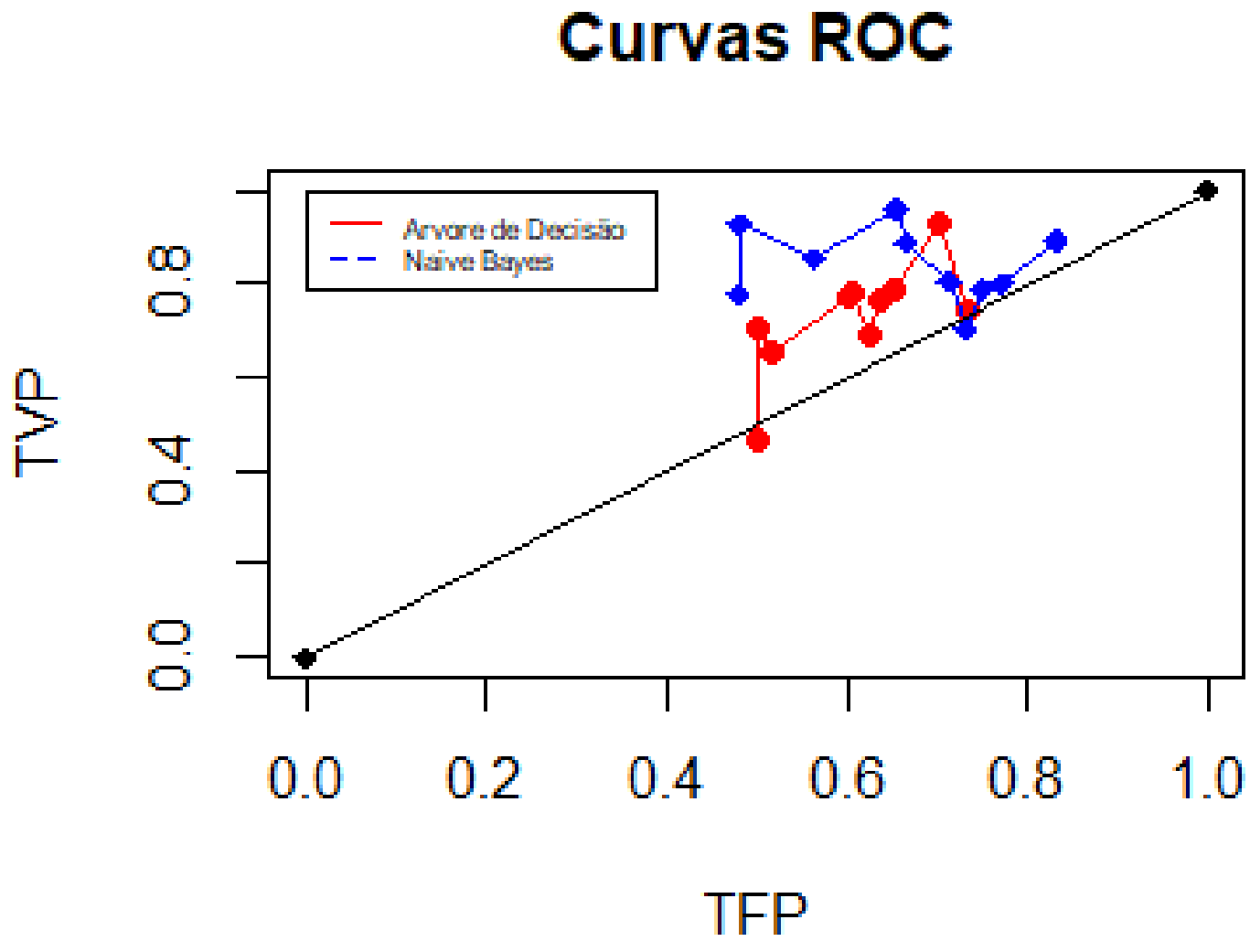
# Validação

- Foi escolhido o método de validação **10-fold cross validation**.
- Método escolhido com base nas discussões em sala.

# Validação

- Foi escolhido o método de validação **10-fold cross validation**.
- Método escolhido com base nas discussões em sala.
- Foram geradas as curvas ROC para cada método com o objetivo de facilitar a visualização dos resultados.

# Curvas ROC


Curvas ROC
Arvore de Decisão
Naïve Bayes
TVP
0.0
0.4
0.8
0.0
0.2
0.4
0.6
0.8
1.0
TFP

# Métricas

Árvore de Decisão

- **Acurácia média: 0.574034**
- **Precisão média: 0.584587**

Naive Bayes

- **Acurácia média: 0.622000**
- **Precisão média: 0.600699**