

# Projeto 3

## (Risco de diabetes em estágio inicial)

Milena Lucas dos Santos  
Lucas Garavaglia

UNIOESTE

28 de julho de 2022

# Conteúdo

- 1 Domínio da aplicação
- 2 Métodos aplicados
- 3 Ferramenta utilizada
- 4 Dificuldades e considerações finais
- 5 Referências

# Domínio da aplicação

## Domínio da aplicação

- Pessoas com diabetes
- Dados reais de um hospital em Sylhet.
- Dados obtidos a partir de um questionário.

# Pré-processamento

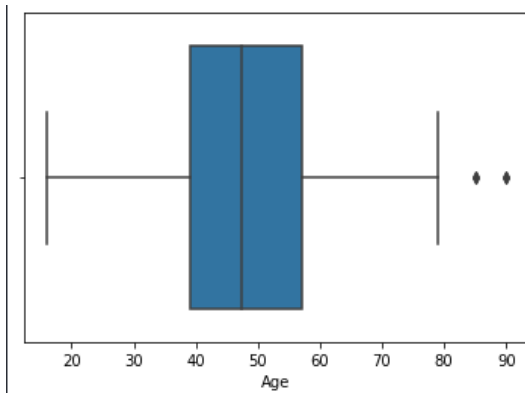
- Valores faltantes.

```
dataset.info()
✓ 0.8s

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 520 entries, 0 to 519
Data columns (total 17 columns):
#   Column                                Non-Null Count  Dtype
---  ---                                ---
0   Age                                  520 non-null    int64
1   Gender                              520 non-null    object
2   Polyuria                             520 non-null    object
3   Polydipsia                           520 non-null    object
4   sudden weight loss                   520 non-null    object
5   weakness                             520 non-null    object
6   Polyphagia                           520 non-null    object
7   Genital thrush                       520 non-null    object
8   visual blurring                      520 non-null    object
9   Itching                              520 non-null    object
10  Irritability                         520 non-null    object
11  delayed healing                      520 non-null    object
12  partial paresis                      520 non-null    object
13  muscle stiffness                     520 non-null    object
14  Alopecia                             520 non-null    object
15  Obesity                              520 non-null    object
16  class                                520 non-null    object
dtypes: int64(1), object(16)
memory usage: 69.2+ KB
```

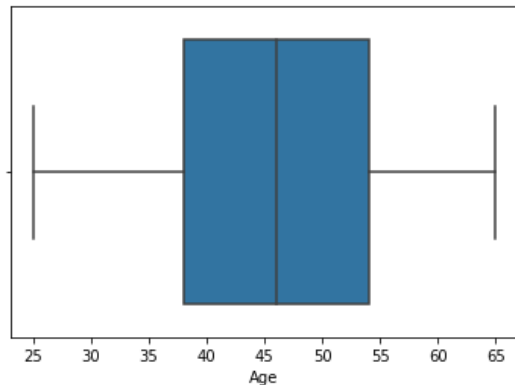
# Pré-processamento

- Detecção de outliers.
- Gráfico box-plot.



# Pré-processamento

- Remoção dos outliers.



# Pré-processamento

- Seleção de atributos.
- Recursive Feature Elimination (RFE).

Selected True, Rank: 1.000	Column: Age
Selected True, Rank: 1.000	Column: Gender
Selected True, Rank: 1.000	Column: Polyuria
Selected True, Rank: 1.000	Column: Polydipsia
Selected False, Rank: 9.000	Column: sudden weight loss
Selected False, Rank: 8.000	Column: weakness
Selected False, Rank: 7.000	Column: Polyphagia
Selected True, Rank: 1.000	Column: Genital thrush
Selected False, Rank: 6.000	Column: visual blurring
Selected False, Rank: 5.000	Column: Itching
Selected True, Rank: 1.000	Column: Irritability
Selected False, Rank: 4.000	Column: delayed healing
Selected False, Rank: 3.000	Column: partial paresis
Selected True, Rank: 1.000	Column: muscle stiffness
Selected True, Rank: 1.000	Column: Alopecia
Selected False, Rank: 2.000	Column: Obesity

# Mineração de dados

- K Nearest Neighbor.
  - ▶ KSimple.
  - ▶ Genérico.
- Árvore de decisão.
  - ▶ Fácil compreensão.
  - ▶ Resultado visual.
  - ▶ Pouco pré processamento.
- Random forest.
  - ▶ Hiperparâmetros.
  - ▶ Soluciona alguns dos problemas da árvore de decisão (overfitting).



# Extração de padrões

- Cross-validation f1-score.

Comparação entre os datasets utilizando KNN:

F1-Score	Descrição
0.84	Dataset com atributos selecionados.
0.82	Dataset com pré-processamento sem atributos selecionados.
0.79	Dataset sem pre-processamento.

Comparação entre os datasets utilizando random florest:

F1-Score	Descrição
0.96	Dataset com atributos selecionados.
0.95	Dataset com pré-processamento sem atributos selecionados.
0.94	Dataset sem pre-processamento.

Comparação entre os datasets utilizando decision tree:

F1-Score	Descrição
0.97	Dataset com atributos selecionados.
0.93	Dataset com pré-processamento sem atributos selecionados.
0.88	Dataset sem pre-processamento.

# Pós-processamento

- Cross-validation f1-score.
- Teste de friedman.
- Teste Friedman nemenyi.

# Pós-processamento

knn	pvalue=0.02472352647033933
Random Florest	pvalue=0.8920030614530929
Árvore de decisão	pvalue=0.3678794411714408

Comparação entre os classificadores:

Algoritmo	F1-Score
-----------	----------

KNN:	0.84
Random Florest:	0.96
Decision Tree:	0.97

# Ferramenta utilizada

- Python.
- Jupyter.
- Sklearn.
- Scipy.

# Dificuldades e considerações finais

- Problema com tipos de dados na biblioteca.

# Referências I

ISLAMEMAIL, M. M. F. et al. O QUE E DIABETES. 2020. Acesso em: 19, abril de 2022. Disponível em:

⟨<https://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset>.⟩

METABOLOGIA, S. brasileira de Endocrinologia e. O QUE E DIABETES. 2007. Acesso em: 19, abril de 2022. Disponível em:

⟨<https://www.endocrino.org.br/o-que-e-diabetes/>⟩.

scikit-learn. 2022. Disponível em: ⟨<https://scikit-learn.org/stable/>⟩.

Scipy. 2022. Disponível em:

⟨<https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.friedmanchisquare.html>.⟩