Breast Cancer Wisconsin (Diagnostic) Data Set

UNIOESTE

GUILHERME BACHEGA GOMES

FOZ DO IGUAÇU, 2022

Errata

- Esta apresentação **teve dois erros identificados e corrigidos após a sua entrega**:

1. Slide número 8: a palavra **DIREÇÃO** estava incorretamente grafada como **DIREAÇÃO**;
2. Slide número 24: o p-valor do algoritmo C4.5 (segunda linha e segunda coluna da tabela) igual a **0.019** estava incorretamente grafado como **0.19**.

A Base de Dados

- É composto por uma série de atributos do núcleo de células do câncer de mama;
- Suas características foram computadas através da digitalização de uma amostra de um seio de uma paciente com câncer de mama;
- Os atributos da base descrevem características do **núcleo**.

Características

- Multivariável;
- 569 registros;
- 32 atributos: 2 de predição (benigno ou maligno) e 30 das características extraídas do núcleo;
- Não há dados faltantes.

Informações sobre os atributos

1. Número de identificação;
2. Diagnóstico (M = maligno, B = benigno);
3. Dez valores reais computados de cada núcleo da célula:
    a) Raio;
    b) Textura;
    c) Perímetro.
    d) Área;
    e) Suavidade;
    f) Compacidade;
    g) Concavidade;
    h) Pontos de concavidade;
    i) Simetria;
    j) Dimensão fractal.

Materiais

- *Google Colab;*
- *Python 3;*
- *Scikit-learn;*
- *Pandas;*
- *Numpy;*
- *Scipy;*
- *Seaborn.*

Pré-processamento

- Remoção do atributo ID;
- Atributo diagnóstico é escolhido como classe;
- ***StandardScaler***
  1. Média próxima de zero e desvio padrão igual a um;
  2. Menos sensível à outliers;
  3. Estimadores do scikit-learn se comportam melhor [7].

Pré-processamento

- **Seleção de atributos**:
  1. *SequentialFeatureSelector* com KNeighbours com K = 5;
  2. Número de atributos selecionados = 10;
  3. Direção = *forward*.

SequentialFeatureSelector

- Método guloso que busca o melhor conjunto de atributos que gera o melhor resultado para o classificador de teste (no caso deste trabalho: *5-nearest neighbors*);
- O parâmetro *forward* o faz começar com nenhum atributo e adiciona novos iterativamente;
- O parâmetro *backward* o faz começar com todos os atributos e os remove iterativamente;
- OBS: há uma grande diferença de resultado na execução do modo *forward* e *backward*.

Pré-processamento

- Após a seleção de atributos, plota-se a matriz de calor da correlação de *Pearson* entre os atributos (demonstrado no *Notebook*);
- Nota-se a existência de atributos fortemente correlacionados (> 90%);
- Opta-se por remover os atributos altamente correlacionados, menos um.

Pré-processamento

- Correlação forte com smoothness_mean:
  1. radius_worst;
  2. perimeter_worst.

- Opta-se por remover *radius_worst* e *perimeter_worst*, mantendo *smoothness_mean*.

Pré-processamento

- 8 atributos finais:
  1. area_mean;
  2. smothness_se;
  3. concavity_se;
  4. symmetry_se;
  5. fractal_dimension_se;
  6. texture_worst;
  7. smothness_worst;
  8. concavity_worst.

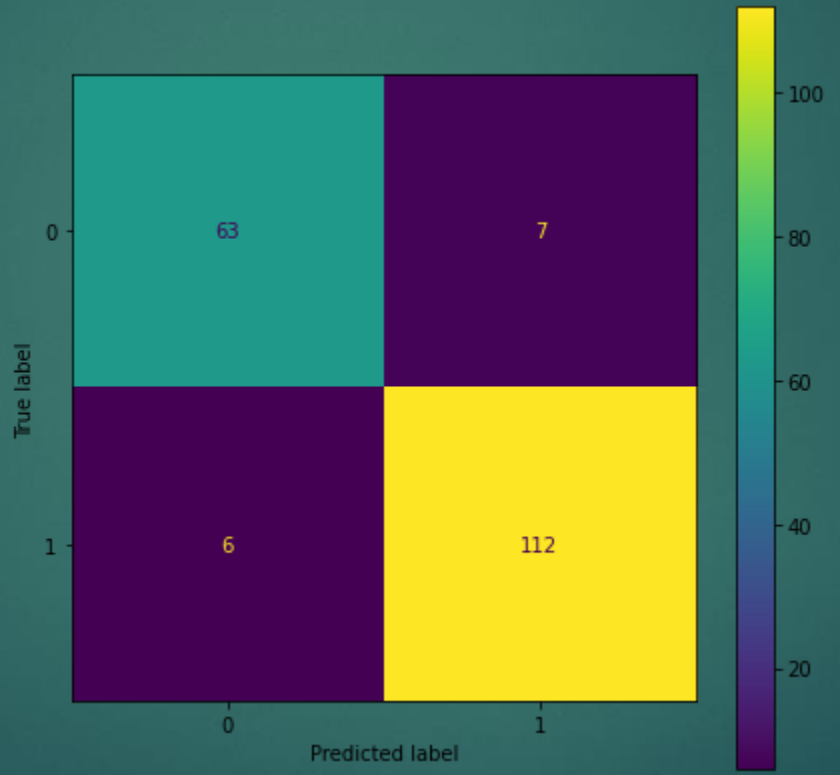
Extração de padrões

- Três classificadores escolhidos: K-nearest neighbors, Árvore de Decisão (C4.5) e *Multilayer Perceptron*;
- Treinados tanto com *stratified holdout* quanto com *10-fold stratified cross-validation*.
- O treino com *stratified holdout* é apresentado por curiosidade, o treino com *cross-validation* é utilizado para a avaliação final dos modelos;
- Sensibilidade escolhida como métrica de desempenho. Por se tratar de diagnóstico de câncer, é muito mais preferível falsos positivos do que falsos negativos.

K-nearest nieghbors (Stratified holdout)

Classe	Precisão	Sensibilidade	*F1-Score*	Instâncias
Maligno	91%	90%	91%	70
Benigno	94%	95%	95%	118

K-nearest nieghbors (Stratified holdout)



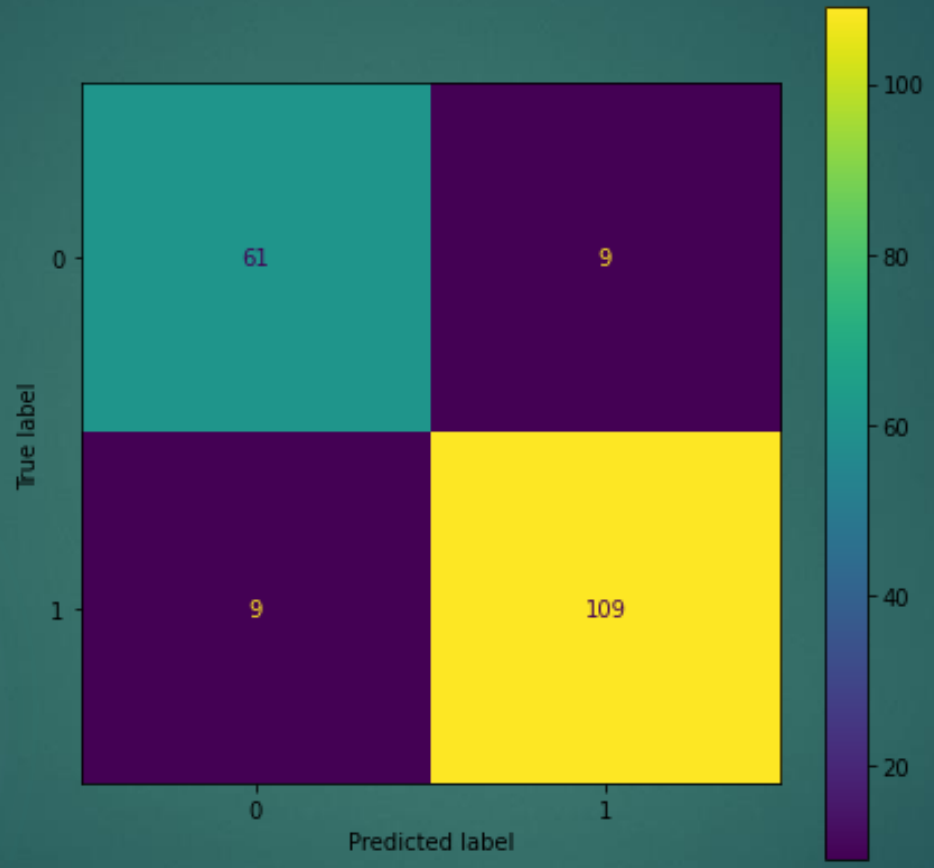
K-nearest nieghbors (10-fold cross-validation)

Fold	Sensibilidade
1	91%
2	88%
3	95%
4	94%
5	97%
6	92%
7	91%
8	96%
9	95%
10	85%
Média	93% +- 3

Árvore de Decisão (C4.5) (*Stratified holdout*)

Classe	Precisão	Sensibilidade	*F1-Score*	Instâncias
Maligno	87%	87%	87%	70
Benigno	92%	92%	92%	118

Árvore de Decisão (C4.5) (*Stratified holdout*)



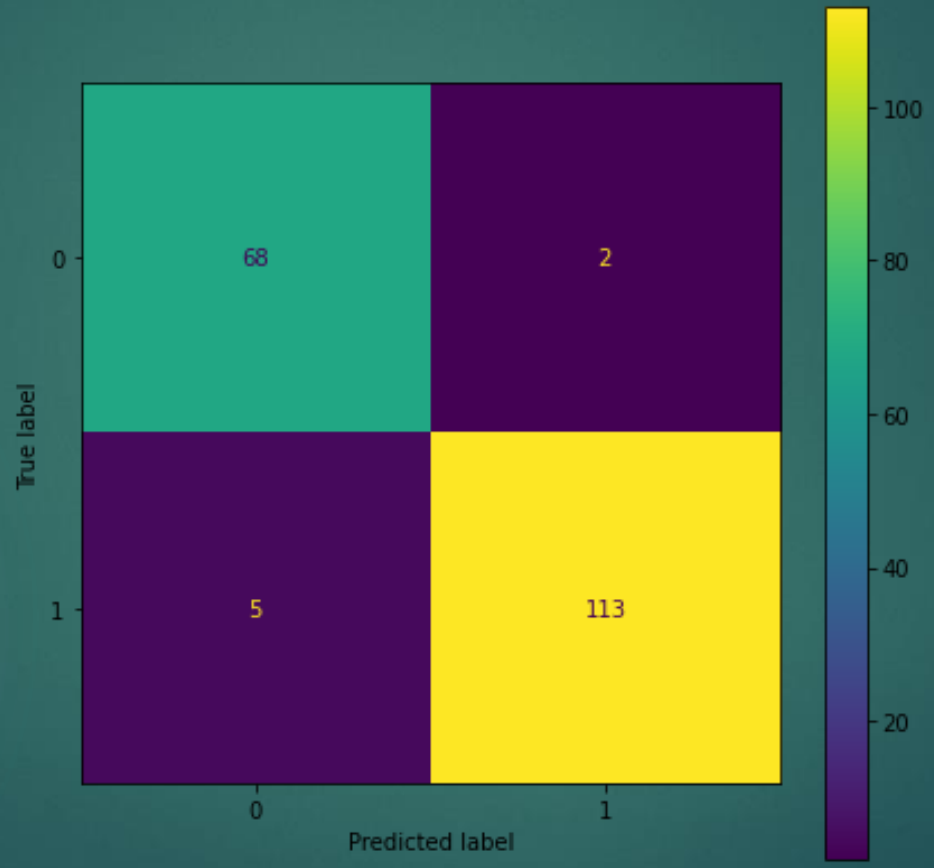
Árvore de Decisão (C4.5) (*10-fold cross-validation*)

Fold	**Sensibilidade**
1	95%
2	86%
3	97%
4	86%
5	97%
6	96%
7	93%
8	97%
9	87%
10	98%
Média	93% +- 4

Multilayer Perceptron (MLP) (*Stratified holdout*)

Classe	Precisão	Sensibilidade	*F1-Score*	Instâncias
Maligno	93%	97%	95%	70
Benigno	98%	96%	97%	118

Multilayer Perceptron (MLP) (Stratified holdout)



Multilayer Perceptron (MLP) (*10-fold cross-validation*)

Fold	**Sensibilidade**
1	98%
2	95%
3	97%
4	97%
5	96%
6	96%
7	91%
8	98%
9	94%
10	92%
Média	95% +- 2

Pós-processamento

- Passos:
  1. Verificar se os dados da sensibilidade de cada *fold* seguem uma distribuição normal;
  2. Verificar se os grupos são pareados (em essência, se os algoritmos foram treinandos com os mesmos *folds*);
  3. Verificar a quantidade de grupos que serão testados.

Teste de Normalidade (*Shapiro-Wilk*)

- Execução de teste Shapiro-Wilk tomando como condição para se rejeitar a hipótese nula (isto é, determinar que os dados não seguem distribuição normal) um p-valor < 0,05.

Statistic	***P-value***
0.93	0.51
0.81	0.019
0.90	0.25

Teste de Normalidade (*Shapiro-Wilk*)

Como o p-valor do grupo gerado pelo algoritmo C4.5 é menor que 0.05 descarta-se os testes paramétricos (mesmo que os outros dois apresentem uma distribuição normal).

Verificar se os grupos são pareados

Pela documentação da função cross_validate da biblioteca scikit-learn é possível determinar que não há o embaralhamento dos folds durante o cross-validation de cada algoritmo. Ou seja: os algoritmos são treinados com os mesmos conjuntos de dados, desta forma **determina-se que os grupos são pareados**.

Verificar a quantidade de grupos que serão submetidos ao teste

A quantidade de algoritmos de aprendizado de máquina executados é igual à três, gerando três grupos de desempenho (sensibilidade em cada *fold*) a serem comparados.

Teste de *Friedman*

- Determinando que os dados não seguem distribuição normal, são pareados e são três grupos, escolhe-se o Teste de *Friedman* para o teste da existência de diferença estatisticamente significativa entre os três grupos. Determina-se a condição para rejeitar a hipótese nula (isto é, haver diferença estatística significativa entre os grupos) um p-valor < 0,05.

Teste de *Friedman*

Statistic	***P-value***
2.5789	0.27

- Conclui-se que **não há diferença estatística significativa entre os grupos e pode-se afirmar que o desempenho das três abordagens é similar**.

Conclusão

- Como não há diferença estatisticamente significativa, recomenda-se o uso do classificador C4.5, pelos seguintes motivos:
  1. Algoritmo caixa-branca, ou seja, é possível entender a linha de raciocínio seguida no processo de classificação;
  2. Menos complexo.

Referências

- [1] BRAMER, M. Principles of Data Mining. [S.l.]: Springer, 2016.
- [2] Breast Cancer Wisconsin (Diagnostic) Data Set. Disponível em https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29. Utilizado a versão em formato CSV disponível em https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data?resource=download. Ambos acessados em 15/07/2022.
- [3] Géron, A. (2019). Hands-on machine learning with Scikit-Learn, Keras and TensorFlow: concepts, tools, and techniques to build intelligent systems (2nd ed.). O'Reilly.

Referências

- [4] Material de aula: Pré-processamento de Dados. Huei Diana Lee. Universidade Estadual do Oeste do Paraná, Foz do Iguaçu.
- [5] Material de aula: Statistical Analysis of Experiments in Data Mining and Computational Intelligence. Salvador García, Francisco Herrera. University of Granada, Spain.
- [6] Tutorial Estatístico - Canal Pesquise. Disponível em https://www.canalpesquise.com.br. Acessado em 28/07/2022.
- [7] Scikit-learn, StandardScaler(). Disponível em https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html. Acessado em 28/07/2022.