Relatório de Aprendizado de Máquina - Aula 5

Lucas Ribeiro da Silva - 2022055564
Universidade Federal de Minas Gerais
Belo Horizonte - Minas Gerais - Brasil

lucasrsilvak@ufmg.br

1 Introdução

Neste relatório, utilizamos a biblioteca SHAP e a Análise de Coeficientes para avaliar a contribuição de cada variável para o resultado final da análise.

2 Dados

Exploração e Preparação dos Dados do Conjunto

Atributo	**0**	**1**	**2**	**3**	**4**
index	0	1	2	3	4
mean radius	17.99	20.57	19.69	11.42	20.29
mean texture	10.38	17.77	21.25	20.38	14.34
mean perimeter	122.8	132.9	130.0	77.58	135.1
mean area	1001.0	1326.0	1203.0	386.1	1297.0
mean smoothness	0.1184	0.08474	0.1096	0.1425	0.1003
mean compactness	0.2776	0.07864	0.1599	0.2839	0.1328
mean concavity	0.3001	0.0869	0.1974	0.2414	0.198
mean concave points	0.1471	0.07017	0.1279	0.1052	0.1043
mean symmetry	0.2419	0.1812	0.2069	0.2597	0.1809
mean fractal dimension	0.07871	0.05667	0.05999	0.09744	0.05883
radius error	1.095	0.5435	0.7456	0.4956	0.7572
texture error	0.9053	0.7339	0.7869	1.156	0.7813
perimeter error	8.589	3.398	4.585	3.445	5.438

Atributo	0	1	2	3	4
area error	153.4	74.08	94.03	27.23	94.44
smoothness error	0.006399	0.005225	0.00615	0.00911	0.01149
compactness error	0.04904	0.01308	0.04006	0.07458	0.02461
concavity error	0.05373	0.0186	0.03832	0.05661	0.05688
concave points error	0.01587	0.0134	0.02058	0.01867	0.01885
symmetry error	0.03003	0.01389	0.0225	0.05963	0.01756
fractal dimension error	0.006193	0.003532	0.004571	0.009208	0.005115

3 Regressão Logística e Curva ROC

Dados da Regressão Logística e Curva ROC

Métrica	Valor
Acurácia	0.9561

Tabela 2: Acurácia do modelo.

Matriz de Confusão	Predição: 0	Predição: 1
Verdadeiro: 0	39	4
Verdadeiro: 1	1	70

Tabela 3: Matriz de confusão do modelo.

Classe	Precisão	Revocação	F1-Score
0	0.97	0.91	0.94
1	0.95	0.99	0.97
Média Acurácia	0.96	0.95	0.95
Média Ponderada	0.96	0.96	0.96

Tabela 4: Relatório de Classificação do modelo.

3.1 Curva ROC

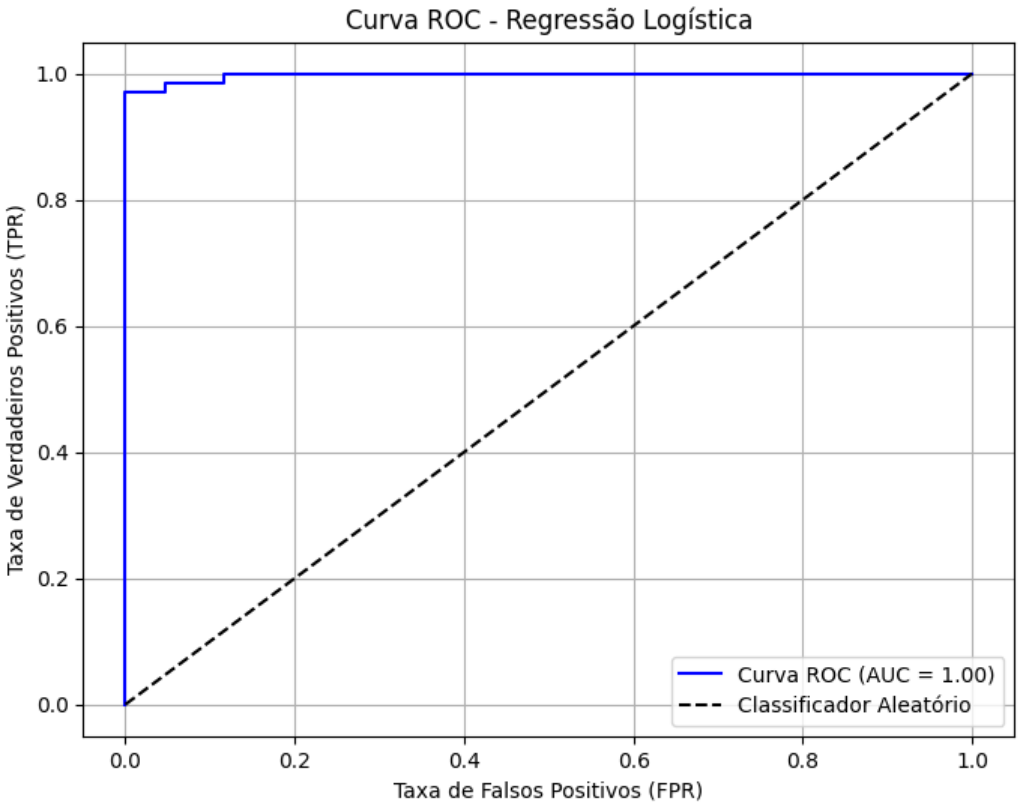


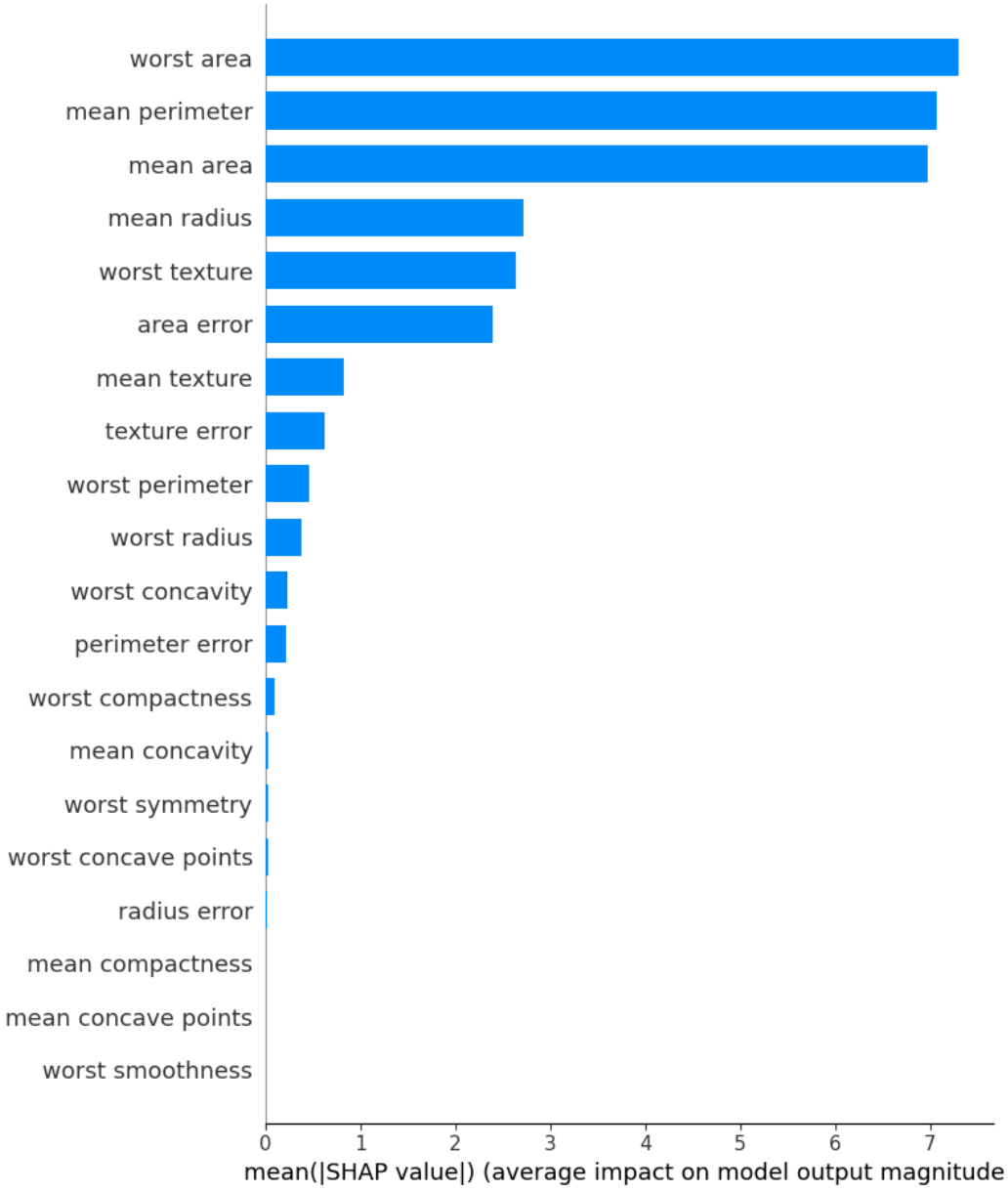
Figura 1: Curva ROC

3.2 Análise

O modelo atingiu um resultado satisfatório com uma acurácia de 95,61% e com uma precisão alta nas duas classes além de uma revocação alta. A curva ROC também apresentou um resultado satisfatório com uma área de praticamente 1 sobre a curva.

4 Análise SHAP

4.1 Análise Geral



De acordo com o gráfico, os itens mais importantes para a previsão da probabilidade de câncer são *worst area*, *mean perimeter*, *mean area*, *mean radius*, *worst texture* e *area error*. Os outros itens contribuem menos ou muito pouco para a interpretação da possibilidade ou não.

4.2 Análise do Item 0

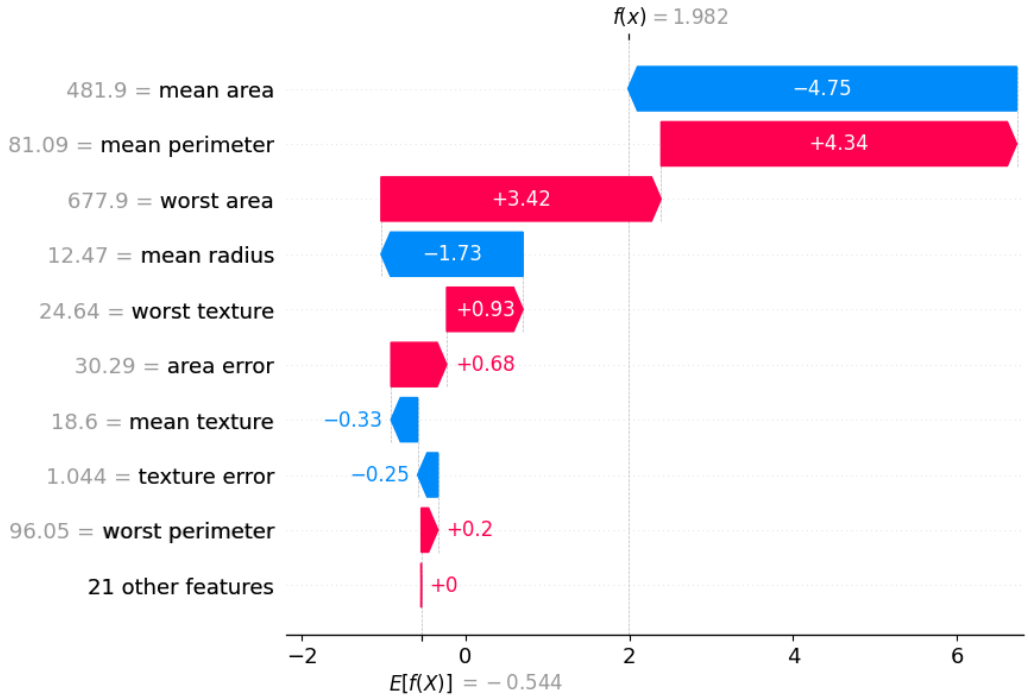


Figura 2: Análise do Item 0

De acordo com as variáveis do Item 0, podemos perceber que a *mean area*, *mean radius* e *mean texture* contribuem muito para a possibilidade de não ter câncer e que *mean perimeter*, *worst area*, *worst texture* e *area error* contribuíam negativamente. A análise geral do Item 0 leva para um $f(x) = 1.982$, onde a probabilidade indica para câncer.

4.3 Análise do Item 1

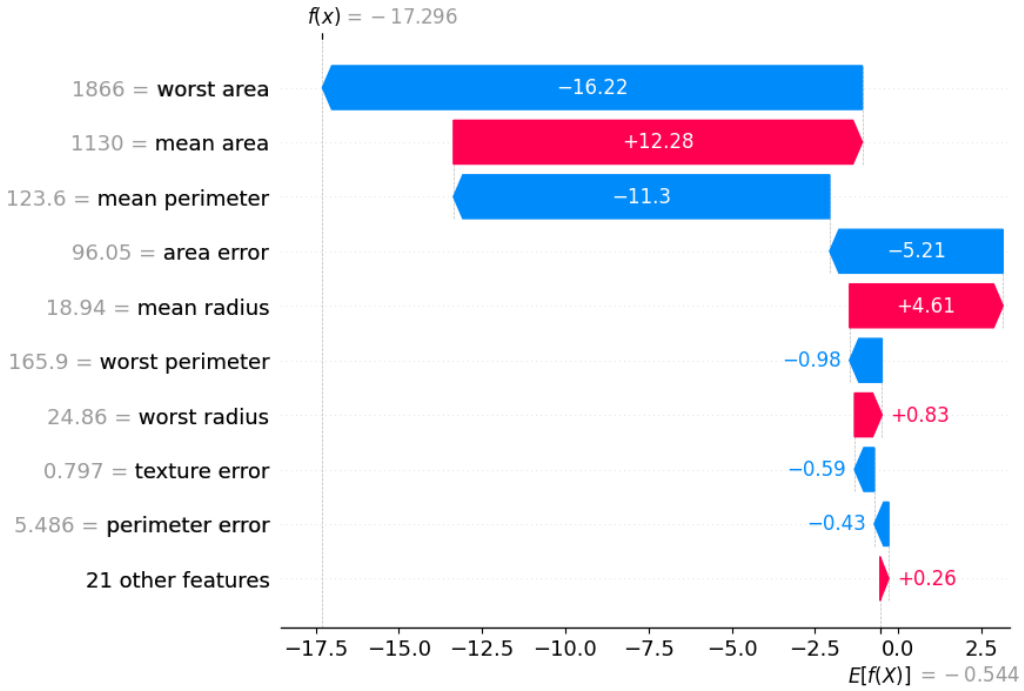


Figura 3: Análise do Item 1

De acordo com as variáveis do Item 1, podemos perceber que a *worst area*, *mean perimeter* e *area error* contribuem muito para a possibilidade de não ter câncer e que *mean area*, *mean radius* e *worst radius* contribuíam negativamente. A análise geral do Item 1 leva para um *f(x) = -17.296*, onde a probabilidade indica para não ter câncer.

4.4 Análise do Item 2

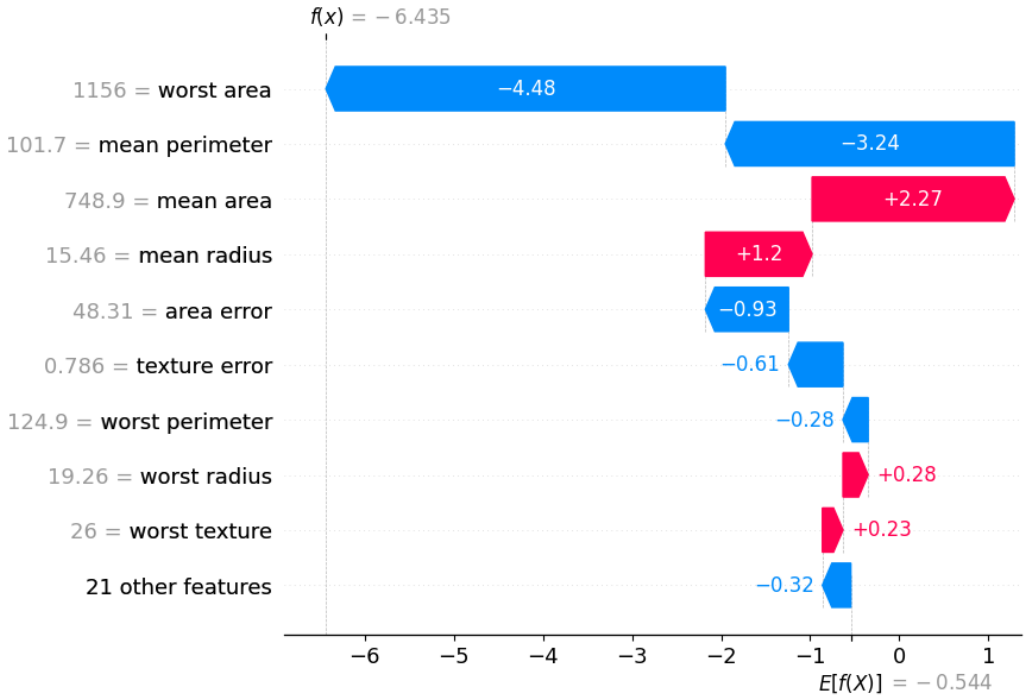


Figura 4: Análise do Item 2

De acordo com as variáveis do Item 2, podemos perceber que a *worst area*, *mean perimeter*, *area error* e *texture error* contribuem muito para a possibilidade de não ter câncer e que *mean area*, *mean radius* e *worst radius* contribuíam negativamente. A análise geral do Item 2 leva para um *f(x) = -6.435*, onde a probabilidade indica para não ter câncer.

4.5 Análise

Parâmetros como *worst area*, *mean perimeter* e *mean area* são significativamente mais importantes para a análise de probabilidades, pela esperança do problema (*E(f(x))* é possível perceber também que o algoritmo tende a começar a classificação de um ponto onde a probabilidade de ter câncer é menor.

5 Conclusão

Ao utilizar o SHAP e fazer a análise de coeficientes, é possível observar a construção de modelos de classificação, avaliação de desempenho e interpretar as variáveis

e observar quais são as variáveis mais importantes na descrição dos resultados bem como permitir a explicação e interpretação dos resultados.