SHAP

Classificação e Interpretação de Variáveis com Regressão Logística e SHAP

1. Exploração e preparação dos dados

Ao carregarmos o conjunto de dados de breast câncer (câncer de mama) e ao visualizarmos os dados, já é perceptível que temos muitas colunas e, consequentemente, **muitas variáveis** para analisar. A partir disso, já podemos imaginar que, sem ferramentas adicionais que ajudem nossa análise, será um pouco difícil compreender os impactos das variáveis no modelo de **Regressão Logística** que queremos fazer, para prever um tumor benigno ou maligno, e é nesse momento que o **SHAP** nos ajuda.

Analisando nossos dados, é perceptível que também vamos ter que fazer uma normalização dos valores, já que possuimos variáveis com valores muito discrepantes entre si, sem estarem dentro de um padrão para o modelo.

Podemos colocar como exemplo a média e o desvio padrão de dois parâmetros:

`mean_area` e `mean_smoothness` :

Medidas / Parâmetros	mean smoothness	mean_area
mean	0.09	654.8
std	0.01	351.9

Nossa variável dependente `diagnosis` possui a seguinte quantidade de amostras para cada classe:

Aa Name	# Number
benigno	357
maligno	212

Portanto, temos mais amostras rotuladas como benigno (1) do que como maligno (0), de um total de 569 amostras.

2. Divisão dos dados

Nesse momento foi feito a divisão dos dados em treino e teste, com o tamanho do conjunto de dados de teste correpondendo a 20% do total dos dados, e é nesse momento em que a normalização dos dados foi realizada, pois a normalização deve ser feita separadamente no conjunto de dados de teste e treino.

É utilizado uma biblioteca pra isso, e o modelo denominado escalador (`scaler`) deve ser ajustado (`fit`) **apenas no conjunto de treino**, e em seguida ser usado para transformar (`transform`) tanto o treino quanto o teste. Isso é essencial para evitar vazamento de informações do conjunto de teste para o treino.

3. Modelo

Agora é construído o modelo de Regressão Logística, realizando o treinamento no nosso conjunto de treino separado para tal fim.

```
model = LogisticRegression()
model.fit(X_train_norm, y_train)
y_pred = model.predict(X_test_norm)

# Calcular métricas importantes
precision = precision_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred)
confusion_matrix = confusion_matrix(y_test, y_pred)
```

```
print(f"Precisão: {precision:.4f}")
print(f"Recall: {recall:.4f}")
print(f"F1-Score: {f1:.4f}")
print(f"Matriz de confusão: \n {confusion_matrix}")
```

As seguintes métricas foram calculadas, com seus respectivos resultados:

```
Precisão: 0.9459
Recall: 0.9859
F1-Score: 0.9655
Matriz de confusão:
 [[39  4]
 [ 1 70]]
Verdadeiros Positivos:  70
Falsos Positivos:  4
Verdadeiros Negativos:  39
Falsos Negativos:  1
```

Lembrando que, os valores de Precisão, Recall e F1-Score funcionam da seguinte forma:

$$\text{Precisão} = \frac{\text{TP}}{\text{TP}+\text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}}$$

$$\text{F1-Sore} = \frac{2 * \text{precisão} * \text{recall}}{\text{precisão}+\text{recall}}$$

Além disso, foi feito a curva ROC, afim de ajudar a analisar a eficiência de nosso modelo também:

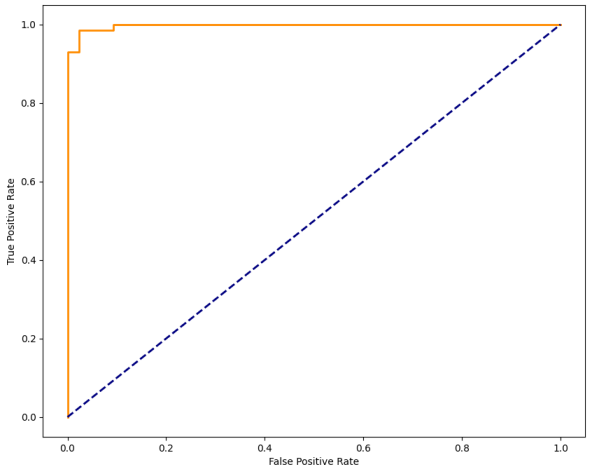

Figura 1 - Curva ROC.

A partir da curva ROC, vemos que a Sensiblidade do modelo, que é a capacidade do modelo de detectar corretamente resultados positivos, está bem alta. Podemos avaliar isso pela Matriz de Confusão também, verificando a quantidade de amostras de verdadeiros positivos (70) e falsos positivos (4).

A especificidade, por sua vez, que é a capacidade do modelo de detectar corretamente resultados negativos (Verdadeiros Negativos) também está próxima de 100%, o que contribui também para a acurácia do modelo.

4. Interpretação do modelo com SHAP values

SHAP values representa o quanto uma variável impacta nas previsões do nosso modelo, seja uma previsão para a classe esperada (1) ou não.

Magnitude e Sinal dos SHAP Values:

Parâmetro com valor negativo: ele "empurra" a probabilidade em direção a classe oposta ou com probabilidade 0.

Parâmetro com valor positivo: ele "empurra" a probabilidade em direção a classe-alvo com probabilidade 1.

Interpretação do gráfico com SHAP values:

- **Impacto no modelo**: a extensão de pontos na horizontal para uma variável indica o alcance e a distribuição dos SHAP values para diferentes instâncias (valores) da variável. Uma extensão mais longa indica uma variável que possui uma magnitude maior de impacto nas previsões do modelo.
- **Sinal dos valores:** Parâmetros com pontos deslocados mais para a direita indica uma contribuição mais positiva (1) para o modelo, enquanto aqueles com pontos deslocados mais para a esquerda indica uma contribuição mais negativa (0).
- **Consistência:** Posições consistentes dos pontos de uma variável para diferentes instâncias indica um impacto estável nas previsões do modelo. Por outro lado, distribuições irregulares pode indicar pouca estabilidade nas previsões do modelo, “variando” a importância dessa variável.

No nosso caso:
Classe 0 representa câncer maligno, e 1 benigno.

Abaixo está representado nossos parâmetros e seus SHAP values de acordo com cada variável:

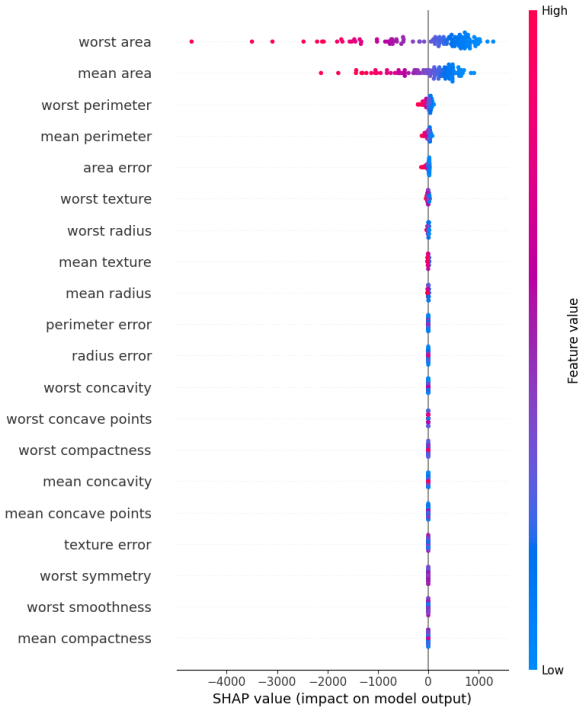

Figura 2.

A partir dos valores obtidos com as variáveis, podemos ver que os parâmetros 'worst_area' e 'mean_area' são os que mais influenciam o modelo a classificar uma amostra como benigna (1) ou maligna (0).

Temos o seguinte significado dessas variáveis:

mean_area:

Descrição: Representa a área média dos núcleos das células nos agregados presentes na amostra.

worst_area:

Descrição: Refere-se à maior (ou "pior") medição da área dos núcleos das células entre os agregados, baseada em métricas como a média de três desvios padrão mais elevados das medições.

Visto que a área e o tamanho de uma célula estão diretamente relacionadas ao tumor ser benigno ou maligno, faz total sentido essas duas variáveis serem as que mais impactam o modelo a classificar uma amostra como 1 ou 0, respectivamente.

Vemos que, a partir da Figura 2, **valores menores** de `mean_area` e `worst_area` impactam o modelo a classificar uma **amostra como benigna** (1), enquanto **valores maiores** dessas variáveis contribuem para uma **classificação maligna.**

5. Ação para campanhas de Marketing

Com base nos resultados obtidos, podemos fazer uma campanha de Marketing voltada a conscientização, prevenção e detecção precoce do câncer de mama.

Campanhas de educação:

Desenvolver campanhas informativas que expliquem como características como **áreas médias e extremas** de lesões podem ser indicativos de malignidade.

Criar materiais educativos, como infográficos e vídeos, para ajudar o público a entender como exames de imagem e medições, como mamografias, podem identificar padrões preocupantes, incentivando check-ups regulares.

Campanhas de detecção:

Para contribuir para a detecção precoce de tumores, é possível realizar exames gratuitos, através do apoio de munícipios e prefeituras, para a análise do tamanho de lesões, principalmente em pessoas que sejam mais propensas a desenvolver o câncer (por meio de genética, por exemplo).

Engajamento com profissionais da área:

Desenvolver workshops ou seminários com base nesses insights para capacitar profissionais de saúde a explicar melhor os resultados de exames aos pacientes.