



# FIAP

# RELATÓRIO TÉCNICO



Alexandre Natã Vicente  
Antônio Cláudio Almeida  
Cyd Ferreira Rodrigues  
David Catherink  
Marcelo Macedo Klotz

**FIAP - Postech em IA para Devs**

## Sumário

<b>1</b>	<b>Sobre o relatório.....</b>	<b>3</b>
<b>2</b>	<b>Introdução .....</b>	<b>4</b>
	2. 1 Links relacionados .....	4
<b>3</b>	<b>Discussões da análise exploratória .....</b>	<b>5</b>
<b>4</b>	<b>Estratégias de pré-processamento. ....</b>	<b>6</b>
<b>5</b>	<b>Modelos utilizados.....</b>	<b>7</b>
<b>6</b>	<b>Resultados e interpretação dos dados .....</b>	<b>8</b>
	6.1 Otimização de Hiperparâmetros.....	11
	6.2 Produção do modelo final e interpretação.....	11
<b>7</b>	<b>Considerações Finais e Aplicabilidade Prática .....</b>	<b>15</b>
<b>8</b>	<b>Atividade extra .....</b>	<b>17</b>

## 1 Sobre o relatório

Este relatório integra o **Tech Challenge** da primeira fase da Pós-Tech em IA para Desenvolvedores (8IADT) da **FIAP**. O projeto consiste no desenvolvimento de uma solução de Inteligência Artificial para o processamento de exames e documentos clínicos. Utilizando fundamentos de **Machine Learning** e **Visão Computacional**, o objetivo central é o auxílio no diagnóstico de câncer de mama, diferenciando tumores malignos de benignos.

O grupo responsável pelo desafio é composto pelos seguintes integrantes da **Secretaria de Segurança Pública do Distrito Federal (SSP/DF)**:

- Alexandre Natã Vicente (**rm370024**) ([ale.n.vicente@gmail.com](mailto:ale.n.vicente@gmail.com))
- Antônio Cláudio Almeida (**rm370052**) ([antonioalmeida@gmail.com](mailto:antonioalmeida@gmail.com))
- Cyd Ferreira Rodrigues (**rm370004**) ([cydnelson@gmail.com](mailto:cydnelson@gmail.com))
- David Catherinck (**rm369997**) ([d.catherinck@gmail.com](mailto:d.catherinck@gmail.com))
- Marcelo Macedo Klotz (**rm370010**) ([marceloklotz@gmail.com](mailto:marceloklotz@gmail.com))

## 2 Introdução

O presente relatório técnico descreve o desenvolvimento de um modelo de Aprendizado de Máquina supervisionado aplicado ao diagnóstico de câncer de mama. O trabalho possui um escopo acadêmico e diversos passos realizados e mantidos para fins acadêmicos.

O banco de dados escolhido foi o [\*Breast Cancer Wisconsin\*](#) (via *Kaggle*), o qual contém atributos morfológicos extraídos de imagens digitalizadas de biópsias de tumores de mama. O problema abordado é de classificação binária, distinguindo tumores benignos e malignos. Neste cenário, erros do tipo falso negativo possuem elevado custo clínico.

### 2.1 Links relacionados

Os arquivos relacionados a este desafio podem ser acessados por meio dos seguintes endereços eletrônicos:

- **Github:** <https://github.com/marceloklotz/breast-cancer-fiap/>
- **Notebook:** [https://github.com/marceloklotz/breast-cancer-fiap/blob/main/FIAP\\_Challenge\\_Breast\\_Cancer.ipynb](https://github.com/marceloklotz/breast-cancer-fiap/blob/main/FIAP_Challenge_Breast_Cancer.ipynb)
- **Base de dados (*Breast Cancer Wisconsin* via *Kaggle*):**  
<https://github.com/marceloklotz/breast-cancer-fiap/blob/main/breast-cancer.csv>
- **Vídeo explicativo:** [https://www.youtube.com/watch?v=8BE0\\_HIziS8](https://www.youtube.com/watch?v=8BE0_HIziS8)

### 3 Discussões da análise exploratória

Na análise exploratório foram utilizadas, dentre outras, as seguintes técnicas: análise das estatísticas descritivas – média, desvio padrão e quartis, histograma de distribuição, matriz de correlação e sua visualização como tabela e mapa de calor, gráficos violin e gráficos de dispersão.

A base de dados é composta por 569 observações e 32 colunas, sendo 30 variáveis numéricas contínuas, uma variável alvo (*diagnosis*) e uma coluna com a identificação da linha – “id”. Não foram identificados valores ausentes, o que assegura consistência estatística e elimina a necessidade de técnicas de imputação.

Observou-se significativa **disparidade de escala entre as variáveis**, com atributos como *area\_mean* apresentando valores de ordem de magnitude muito superiores a variáveis como *fractal\_dimension\_mean* (valores máximos respectivos: 2501 e 0,1).

Assim, se torna relevante aplicação de padronização, especialmente para modelos sensíveis à escala, como o KNN e Redes Neuras e para aplicação de PCA para redução de dimensionalidade.

A variável de interesse, *diagnosis*, é uma variável categórica, com classes Maligno (M) e Benigno (B), que foram codificadas respectivamente para 1 e 0. Observou-se um **leve desbalanceamento entre as classes**, com aproximadamente 62,7% de tumores benignos e 37,3% de tumores malignos.

Da análise das correlações, verificou-se que:

- a) características como *concave points\_worst* (pior pontos côncavos) e *perimeter\_worst* (pior perímetro) mostram que tumores malignos tendem a ser significativamente maiores e irregulares.
- b) há a altíssima correlação, próximo de 1, entre as features de tamanho (*radius*, *perimeter*, *area*) nas suas versões *mean*, *se* e *worst*. Modelos lineares podem ser afetados por esta multicolinearidade.

## 4 Estratégias de pré-processamento.

Diante das características observadas na Análise Exploratória de Dados (EDA), foram adotadas as seguintes estratégias de pré-processamento:

a) descartada a coluna *id*.

b) codificação da variável categórica alvo (*diagnosis*) em Maligno (M) para 1 e Benigno (B) para 0.

c) padronização das variáveis por meio do *StandardScaler*, transformando cada atributo segundo a equação  $z = (x - \mu) / \sigma$ , ou seja, tornando todos com média zero e desvio padrão 1. **A padronização das variáveis foi aplicada apenas no conjunto de treino**, evitando-se o chamado *data leak* (uso de informação do conjunto de teste no treinamento do modelo).

d) aplicação da Análise de Componentes Principais (PCA) a fim de reduzir a dimensionalidade. Utilizou-se 95% da variância explicada. O PCA também foi aplicado apenas ao conjunto de treino, evitando-se o *data leak*. O PCA permite **reduzir a dimensionalidade do espaço de atributos**, mitigar redundâncias e **melhorar a estabilidade numérica do modelo, especialmente diante de alta multicolinearidade**, preservando a maior parte da informação relevante.

Todo o fluxo foi encapsulado em um *pipeline*, prevenindo vazamento de dados e assegurando reprodutibilidade.

No notebook *jupyter*, são mostrados 3 blocos de treino e avaliação dos modelos (5, 6 e 7). O bloco 5, realiza treinamento e teste apenas com divisão treino/teste (*hold-out*), no qual não foi aplicado PCA. O bloco 6, realiza validação cruzada estratificada (*K-Fold*,  $k=10$ ), no qual também não foi aplicado o PCA. O bloco 7, realiza a validação cruzada estratificada com PCA, mantendo 95% da variância explicada.

Tal procedimento foi realizado para fins didáticos, com o objetivo de verificar o comportamento dos modelos com *hold-out* vs *K-Fold* e analisar o impacto do uso do PCA.

## 5 Modelos utilizados

Foram comparados os seguintes algoritmos: Regressão Logística, *K-Nearest Neighbors (KNN)*, *Random Forest*, *Naive Bayes*, *XGBoost*, *Rede Neural* e *SVM*. Foi feita uma seleção contendo modelos apresentados no curso, lineares e não-lineares, simples e complexos, de baixo e alto custo computacional. O intuito foi verificar o desempenho dessa gama diversificada de algoritmos e observar se alguma classe ou algum algoritmo específico se mostraria muito superior.

## 6 Resultados e interpretação dos dados

Os modelos foram treinados de três formas, conforme mostrado nos itens 5, 6 e 7 do *notebook jupyter*:

- O bloco 5, realiza treinamento e teste apenas com divisão treino/teste (hold-out). Foi utilizado 80% da amostra pra treino e 20% da amostra para teste;
- O bloco 6, realiza validação cruzada estratificada com *K-Fold*. Foi utilizado o  $k=10$ ;
- O bloco 7, realiza a validação cruzada estratificada com *K-Fold*,  $k=10$ , e aplicado o PCA, mantendo 95% da variância explicada.

Os resultados obtidos encontram-se nas figuras 1 a 3 mostradas abaixo.

	Accuracy	Precision	Recall	F1-score	Balanced_Accuracy
<b>LogisticRegression</b>	0.982456	1.000000	0.952381	0.975610	0.976190
<b>NeuralNetwork (limiar 0.4)</b>	0.982456	1.000000	0.952381	0.975610	0.976190
<b>SVM</b>	0.982456	1.000000	0.952381	0.975610	0.976190
<b>XGBoost</b>	0.973684	1.000000	0.928571	0.962963	0.964286
<b>RandomForest</b>	0.973684	1.000000	0.928571	0.962963	0.964286
<b>KNeighborsClassifier (k=5)</b>	0.956140	0.974359	0.904762	0.938272	0.945437
<b>NaiveBayes</b>	0.921053	0.923077	0.857143	0.888889	0.907738

Figura 1 – Desempenho dos modelos no Hold-Out.



K-fold Cross-Validation Results:

	Accuracy	Precision	Recall	F1-score	Balanced_Accuracy
SVM_CV	0.978947	0.977866	0.966667	0.971378	0.976349
NeuralNetwork_CV (limiar 0.4)	0.975439	0.973715	0.962121	0.966571	0.972648
LogisticRegression_CV	0.978947	0.991107	0.952597	0.970425	0.973481
XGBoost_CV	0.957863	0.955178	0.933983	0.942637	0.952983
RandomForest_CV	0.956140	0.949114	0.933550	0.939783	0.951378
KNeighborsClassifier_CV (k=5)	0.970144	0.990455	0.929004	0.957883	0.961724
NaiveBayes_CV	0.931548	0.925040	0.895671	0.906102	0.924145

Figura 2 – Desempenho dos modelos utilizando Validação Cruzada tipo K-Fold, k=10.

K-fold Cross-Validation + PCA Results:

	Accuracy	Precision	Recall	F1-score	Balanced_Accuracy
LogisticRegression_CV_PCA	0.980702	0.986561	0.962121	0.973336	0.976815
NeuralNetwork_CV_PCA (limiar 0.4)	0.977193	0.978261	0.962121	0.968995	0.973997
SVM_CV_PCA	0.977162	0.977866	0.961905	0.968939	0.973968
XGBoost_CV_PCA	0.963158	0.964501	0.938095	0.949050	0.957817
RandomForest_CV_PCA	0.947368	0.936265	0.928788	0.929514	0.943481
KNeighborsClassifier_CV_PCA (k=5)	0.963127	0.977955	0.924242	0.948356	0.955137
NaiveBayes_CV_PCA	0.917481	0.914492	0.862121	0.884150	0.905981

Figura 3 – Desempenho dos modelos utilizando PCA (95%) e Validação Cruzada tipo K-Fold, k=10.

Considerando o contexto médico do problema, a métrica priorizada foi o Recall<sup>1</sup> (Sensibilidade), uma vez que o erro de maior impacto clínico é o Falso Negativo, isto é, classificar um tumor maligno como benigno.

Os resultados experimentais indicaram que:

- **Regressão Logística:** com PCA e no *Hold-Out* apresentou os maiores valores de *Recall* e *F1-Score* entre os modelos avaliados, demonstrando elevada capacidade de identificar corretamente tumores malignos;

<sup>1</sup> Recall = TP / (TP + FN)

- **SVM:** apresentou o melhor *Recall* no *Hold-Out*, na validação cruzada e o terceiro melhor *Recall* com PCA, com desempenho muito próximo ao da Regressão Logística, porém com menor interpretabilidade;
- **Redes Neurais:** alcançou bons resultados ( $\text{Recall} > 0,95$  e  $\text{F1 – Score} > 0,95$ ), mas à custa de maior complexidade e menor transparência no processo decisório;
- **XGBoost:** alcançou bons resultados ( $\text{Recall} > 0,92$  e  $\text{F1 – Score} > 0,94$ );
- **Random Forest:** alcançou bons resultados ( $\text{Recall} > 0,92$  e  $\text{F1 – Score} > 0,92$ ); e
- **KNN e Naive Bayes:** apresentaram desempenho inferior, especialmente em *Recall*, tornando-os menos adequados para aplicações médicas sensíveis ao erro.

A exceção do *Naive Bayes*, que apresentou *Recall* em todos os cenários inferior a 0,9, e o KNN, que apresentou no cenário *Hold-out Recall* próximo de 0,9, os demais modelos apresentaram bons resultados e, em certa medida, pode-se considerar bem próximos ou equivalentes.

Para fins de escolha, elegeu-se a Regressão Logística com aplicação do PCA para modelo de produção pelos seguintes motivos:

- A Regressão Logística manteve desempenho elevado e estável nos três cenários avaliados, com destaque para os resultados obtidos quando combinada com PCA. Ou seja, possui um alto poder preditivo;
- A multicolinearidade das features é tratada com o PCA, mantendo a estabilidade do modelo;
- Há redução de atributos e menor custo computacional, uma vez que o PCA e a Regressão Logística são transformações lineares, portanto, bastante simples computacionalmente de aplicar com o modelo treinado (e até mesmo para o treinamento);
- A interpretabilidade global (algo bastante desejado na área de saúde) é possível com a combinação de PCA com a Regressão Logística.

## 6.1 Otimização de Hiperparâmetros

A Regressão Logística modela a probabilidade condicional da classe positiva segundo a função sigmoide:

$$P(y = 1 \mid x) = 1 / (1 + e^{\{-(\beta_0 + \beta_1 x_1 + \dots + \beta_n x_n)\}})$$

O modelo de Regressão Logística da biblioteca *scikit-learn* (*LogisticRegression*) possui diversos parâmetros, tais como: *C* (força da regularização), *class\_weight* (ponderação das classes), *max\_iter* (número máximo de iterações), *tol* (tolerância de convergência), *penalty* (termo de regularização) e *solver* (algoritmo de otimização).

Foram estabelecidos os seguintes parâmetros para o modelo: *max\_iter*=1000000, *tol* = 1e-9, *penalty* = 'l2', *solver* = 'lbfgs'. O *max\_iter*, *tol* e o *solver* foram escolhidos por serem razoáveis para uma boa convergência. O *penalty* l2 foi escolhido por manter todas as variáveis no modelo.

Para os parâmetros *C* e *class\_weight* foi realizada a otimização de hiperparâmetros por meio do *GridSearchCV*. Como resultado, foi obtido *C*= 0,23 e *Class\_weight* = 'balanced'.

## 6.2 Produção do modelo final e interpretação

Para a produção do modelo final, foi utilizado o seguinte *pipeline*:

1. Padronização dos dados (*StandardScaler*);
2. Redução de dimensionalidade (PCA, 95%);
3. Classificação com Regressão Logística otimizada.

No treinamento deste modelo final foi utilizado **todo o conjunto de dados disponível** garantindo que o modelo aproveite ao máximo as informações presentes na base. Isto não é um erro teórico, pois não se está mais avaliando o modelo, e sim, a prática recomendada, pois não há razão em desperdiçar informações disponíveis.

Embora o PCA reduza a dimensionalidade dos dados, **a interpretação dos coeficientes do modelo continua sendo possível\*** pois o PCA é uma **transformação linear**.

A Regressão Logística é ajustada sobre as componentes principais. No entanto, como cada componente é uma combinação linear das variáveis originais, é possível **reconstruir os coeficientes no espaço original**. Para isto, basta utilizar a seguinte equação:

$$\beta_{\text{original}} = \text{components}^T \beta_{\text{pca}}$$

Em que:

- *components* é a matriz de componentes principais;
- $\beta_{\text{pca}}$  são os coeficientes da Regressão Logística no espaço do PCA;
- $\beta_{\text{original}}$  são os coeficientes da Regressão Logística no espaço original.

A figura 4 mostra os nove maiores valores absolutos dos coeficientes beta da Regressão Logística final.

	feature	coeficiente	coef_abs
21	texture_worst	0.715859	0.715859
13	area_se	0.673979	0.673979
24	smoothness_worst	0.660349	0.660349
1	texture_mean	0.626664	0.626664
23	area_worst	0.616844	0.616844
10	radius_se	0.607848	0.607848
20	radius_worst	0.587137	0.587137
22	perimeter_worst	0.577899	0.577899
19	fractal_dimension_se	-0.575800	0.575800

Figura 4 – Valores dos coeficientes beta da Regressão Logística final.

Para a interpretação dos coeficientes, deve-se lembrar que **coeficientes positivos aumentam a probabilidade de classificação como maligno**. Por sua vez, coeficientes negativos aumentam a probabilidade de classificação como benigno. Quanto maior o valor absoluto, maior a influência da feature.

Assim, as variáveis mais relevantes foram: *texture\_worst*, *area\_se* e *smoothness\_worst*.

**Essas features estão associadas a textura e área, conforme esperado, pois quanto mais irregular e maior o tumor, mais chance de ele ser maligno. Desta forma, verifica-se que o modelo utiliza corretamente os preditores, reforçando o que já se conhece de tumores.**

Apesar do uso de SHAP não ser o mais usual para Regressão Logística (pois esta já possui coeficientes diretamente interpretáveis), foi utilizado por fins didáticos, de representação visual e de comparação. Facilitando, assim, a análise e a comunicação dos resultados.

Devido à presença do PCA no pipeline, foi necessário realizar **adaptações em relação ao uso padrão do SHAP**. Em particular, os coeficientes do modelo foram reconstruídos no espaço das variáveis originais, explorando o fato de que tanto o PCA quanto a Regressão Logística são transformações lineares. Essa reconstrução permitiu aplicar o SHAP sobre um modelo linear equivalente no espaço original das features.

Os gráficos SHAP tornam visualmente evidente a distinção entre variáveis com **coeficientes positivos e negativos**. Observa-se na figura 5 que, para a maioria das variáveis, valores mais elevados estão associados a uma maior probabilidade de classificação como maligno, o que é consistente com o conhecimento clínico do problema. **Tumores malignos tendem a apresentar maior crescimento, maior área e maior irregularidade, refletindo-se diretamente nas variáveis analisadas.**

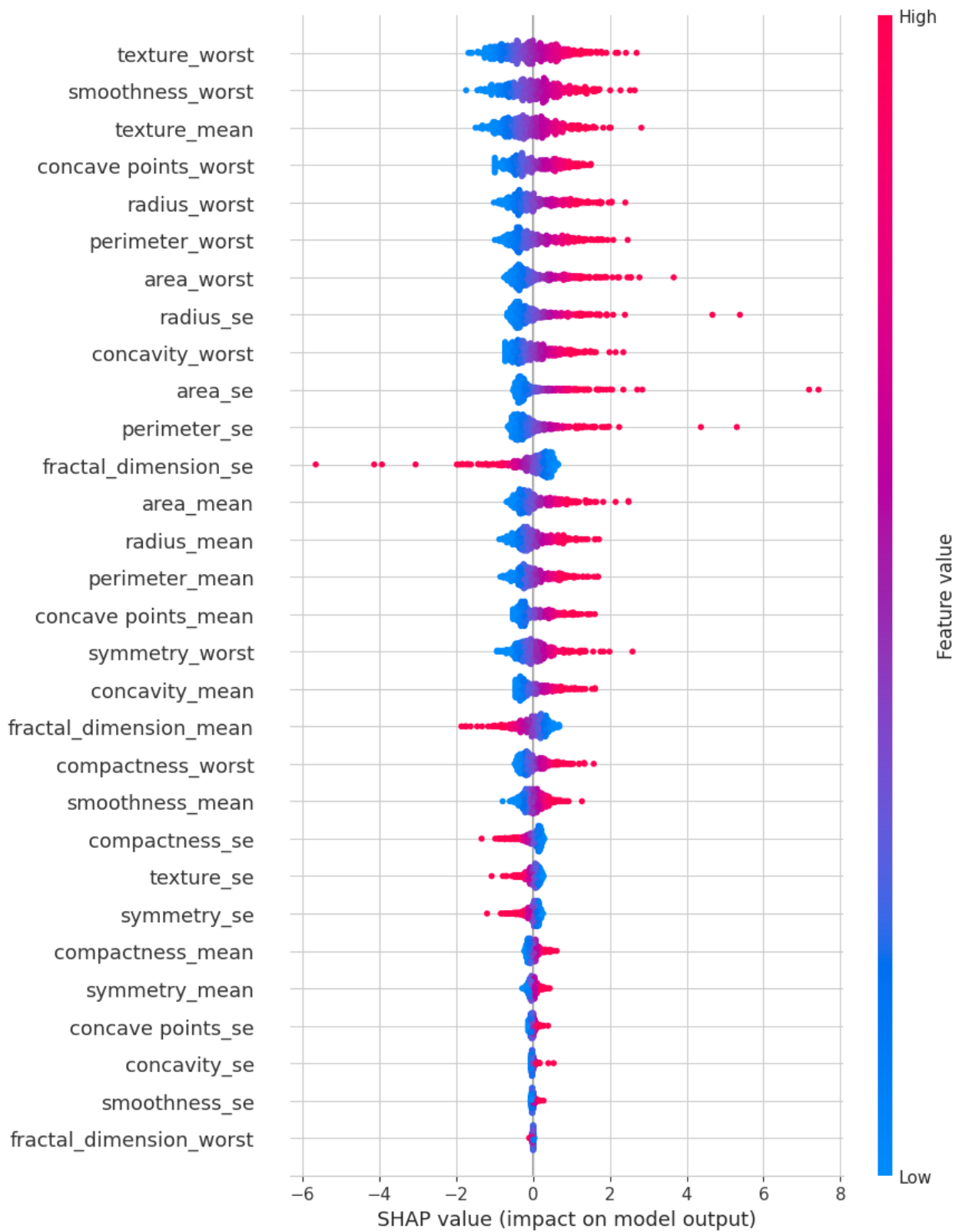


Figura 5 – SHAP do modelo de Regressão Logística final.

## 7 Considerações Finais e Aplicabilidade Prática

Este trabalho teve como objetivo desenvolver e avaliar um modelo de classificação para diagnóstico de câncer de mama, utilizando técnicas de aprendizado de máquina aliadas a boas práticas de pré-processamento, validação e interpretabilidade.

A solução adotada combinou padronização dos dados, redução de dimensionalidade via PCA e Regressão Logística com hiperparâmetros otimizados, resultando em um modelo estatisticamente consistente, computacionalmente eficiente e adequado ao contexto médico analisado.

A Regressão Logística foi escolhida por sua simplicidade, robustez e capacidade de interpretação.

O uso do PCA permitiu reduzir a dimensionalidade do problema, minimizando redundâncias e multicolinearidade entre variáveis, além de contribuir para a estabilidade do modelo.

Apesar da aplicação do PCA, foi possível recuperar a interpretabilidade das variáveis originais por meio da reconstrução dos coeficientes, explorando o fato de que o PCA é uma transformação linear. Isso possibilitou identificar quais características clínicas mais influenciam a classificação.

O modelo final pode ser aplicado como **ferramenta de apoio à decisão**, auxiliando profissionais de saúde na análise de exames e na identificação de casos com maior probabilidade de malignidade.

**Suas principais vantagens práticas incluem:**

- Baixo custo computacional;
- Facilidade de integração em sistemas existentes;
- Respostas rápidas para novos dados;
- Capacidade de explicar o comportamento do modelo, fator essencial em ambientes clínicos.

Os resultados obtidos demonstram que o modelo proposto é tecnicamente sólido e apresenta elevado potencial como ferramenta de apoio à decisão médica. Ressalta-se, entretanto, que sua utilização deve ocorrer exclusivamente como suporte ao diagnóstico, cabendo ao profissional de saúde a decisão final.

Entre as limitações do estudo, destacam-se o uso de um único banco de dados e a ausência de validação externa em dados clínicos reais. Estudos futuros podem incorporar novos conjuntos de dados, modelos mais complexos e técnicas adicionais de explicabilidade, ampliando a robustez e a generalização da solução.

Em síntese, o trabalho atende plenamente aos requisitos do desafio proposto, demonstrando aplicação consistente de fundamentos de *Machine Learning*, análise crítica dos resultados e alinhamento com princípios éticos e clínicos.



## 8 Atividade extra

A seguir link para atividade extra (notebook) realizada pelo grupo com uso de redes neurais convolucionais (CNNs) para o diagnóstico a partir de imagens médicas:

- [https://github.com/marceloklotz/breast-cancer-fiap/blob/main/Atividade\\_Extra.ipynb](https://github.com/marceloklotz/breast-cancer-fiap/blob/main/Atividade_Extra.ipynb)

The image features a solid red background. In the center, the letters "FIAP" are written in a white, modern, sans-serif font. A thin, light red curved line starts from the top right and sweeps across the middle of the image, passing behind the text. At the bottom, there is a horizontal row of five small, light red dots. On the right side, there is a faint, light red curved shape that resembles a stylized arch or a partial circle.

FIAP