UNIVERSIDADE PRESBITERIANA MACKENZIE

Faculdade de Computação e Informática

Ciência De Dados

Projeto Aplicado I – UW Health University Hospital

Gabriel Chaves Gonçalves

São Paulo

2024

Sumário

Lista de Figuras

1. NOMENCLATURA

Dataset: conjunto de dados

Float: dados que representam números com partes decimais

Int: dados que representam números inteiros, que são números sem partes fracionárias

Object: classificação de dados utilizada quando os dados não se enquadram em tipos de dados específicos como int, float,ou bool.

Target: variável alvo

Citologia: ramo da Biologia que estuda as células, suas estruturas e funções, e que pode ser usada p.ara diagnosticar ou rastrear doenças.

AED: análise exploratória de dados

Overfitting: ocorre quando um modelo de aprendizado de máquina aprende muito bem os detalhes e ruídos específicos do conjunto de treinamento, o que o torna menos eficaz em dados novos

2. INTRODUÇÃO

A ciência de dados tem desempenhado um papel fundamental na área da saúde, trazendo avanços significativos na análise de grandes volumes de dados e na criação de modelos preditivos que auxiliam na tomada de decisões médicas. Através da análise de dados complexos, como registros médicos, exames e dados genômicos, os profissionais de saúde podem identificar padrões e fatores de risco, personalizando tratamentos e melhorando o diagnóstico de doenças. Segundo Consoli, *et al*. (2019), no livro *Data Science for Healthcare*, a aplicação de técnicas de aprendizado de máquina e análise preditiva tem revolucionado a medicina, possibilitando a detecção precoce de doenças e a redução de custos operacionais. Além disso, Davenport e Kalakota (2019), em *The Potential for Artificial Intelligence in Healthcare*, destacam como o uso de inteligência artificial e ciência de dados está proporcionando diagnósticos mais precisos e planos de tratamento personalizados.

Nesse contexto, o presente estudo tem como objetivo principal a análise e predição da probabilidade de tumores serem malignos ou benignos, utilizando dados provenientes do *Breast Cancer Wisconsin Dataset*. Com base nas características fisiológicas extraídas de exames médicos, como a

concavidade, o raio e a área dos tumores, aplicamos métodos de aprendizado de máquina, em especial a regressão logística, para identificar os padrões que mais influenciam a malignidade das lesões. A escolha dessa técnica de classificação deve-se à sua capacidade de modelar a relação entre variáveis independentes e a variável-alvo (diagnóstico), possibilitando não apenas a categorização dos tumores, mas também o cálculo de probabilidades. O trabalho inclui uma abordagem completa, desde a análise exploratória dos dados, passando pela seleção das características mais relevantes, até a construção de um modelo preditivo, oferecendo insights importantes para a identificação precoce do câncer de mama.

3. TIPO DE ORGANIZAÇÃO

Fundado em 1924, *UW Health University Hospital* possui uma rica história de atendimento à comunidade e de educação em saúde. Ao longo das décadas, o hospital evoluiu para se tornar uma das principais instituições de saúde tanto no estado quanto nos Estados Unidos da América.

Afiliado à *University of Wisconsin-Madison*, localizado na extremidade oeste do campus, o hospital se beneficia da associação com uma das mais renomadas universidades de pesquisa dos Estados Unidos. Essa relação estreita permite uma integração harmoniosa entre pesquisa, educação e atendimento ao paciente, promovendo inovações significativas nos cuidados de saúde.

Com 614 leitos e 127 clínicas ambulatoriais, o hospital oferece uma ampla gama de serviços médicos e cirúrgicos, apoiado por uma infraestrutura moderna e equipes multidisciplinares. O *UW Health* atende mais de 800.000 pacientes por ano e emprega mais de 24.000 funcionários em sete hospitais, quatro centros médicos, mais de 90 clínicas especializadas, três unidades de cuidados de urgência, quatro centros de saúde comportamental e um centro de serviços comunitários. Comprometido com o cuidado integral dos pacientes, o hospital proporciona tratamentos avançados e suporte contínuo durante todo o processo de recuperação.

- **Missão**

Avançar a saúde sem compromissos por meio de serviço, aprendizado, ciência e responsabilidade social.

- **Visão**

Cuidados de saúde notáveis.

- **Valores**
  - Excelência e Inovação
  - Compaixão
  - Integridade
  - Respeito
  - Responsabilidade e Diversidade

Em relação à participação de mercado, o *UW Health University* Hospital ocupa uma posição significativa no cenário de saúde de Wisconsin. Recentes relatórios o reconhecem como um dos principais hospitais do estado, oferecendo uma ampla gama de serviços especializados, como tratamento de câncer, serviços de transplante, cuidados cardíacos avançados e foco em pesquisa e inovação.

4. CONTEXTO DO ESTUDO

Para aprimorar a precisão no diagnóstico e a eficácia dos tratamentos relacionados a oncologia, o hospital está utilizando um conjunto de dados conhecido como o "*Breast Cancer Wisconsin Dataset*". Este conjunto de dados é amplamente utilizado em estudos de câncer de mama e contém informações detalhadas sobre características de tumores, que são cruciais para a avaliação e tomada de decisão clínica.

O *dataset Breast Cancer Wisconsin* inclui variáveis que ajudam a classificar os tumores como benignos ou malignos com base em medições específicas, como a textura e o tamanho das células. Utilizando essas informações, a equipe de oncologia pode treinar modelos preditivos e realizar análises que auxiliam na detecção precoce de cânceres e na escolha das melhores estratégias de tratamento.

O uso deste conjunto de dados permite ao hospital integrar práticas baseadas em dados ao atendimento clínico, promovendo um ambiente de cuidados mais eficiente e adaptado às necessidades de cada paciente. A análise dos dados proporciona insights valiosos que ajudam a personalizar tratamentos e melhorar os desfechos para os pacientes.

5. DESCRIÇÃO DE ORIGEM

O *Breast Cancer Wisconsin Dataset* tem sua origem em um estudo conduzido pelo Dr. William H. Wolberg, da Universidade de Wisconsin-Madison. O conjunto de dados por meio de uma colaboração entre médicos e pesquisadores com o objetivo de criar um recurso valioso para a análise e a pesquisa no campo da oncologia. Os dados foram obtidos a partir de exames em pacientes com suspeita de câncer de mama, exames de mamografia e biópsias realizadas em pacientes com suspeita de câncer de mama.

A coleta dos dados ocorreu em um contexto clínico, onde as características dos tumores foram registradas para fornecer informações detalhadas sobre suas propriedades. O propósito era construir um banco de dados que pudesse ser utilizado para desenvolver e validar modelos preditivos para a classificação de tumores e melhorar a precisão dos diagnósticos.

6. OBJETIVO DO ESTUDO

O objetivo principal deste estudo é realizar a predição de casos de câncer, utilizando o algoritmo de regressão logística. Através dessa abordagem, buscamos identificar e analisar as características que mais influenciam a variável alvo, que indica se os tumores são malignos ou benignos.

7. CRONOGRAMA

- **Etapa 1:** metas e milestones – Data 02/09/2024
- **Etapa 2:** definição do produto – Data 30/09/2024
- **Etapa 3:** storytelling – Data 28/10/2024
- **Etapa 4:** encerramento – Data 25/11/2024

8. PROPOSTA ANALÍTICA

A presente investigação tem como objetivo realizar a predição da probabilidade de tumores serem malignos ou benignos, utilizando dados do Breast Cancer Wisconsin Dataset. Para alcançar esse objetivo, estou adotando uma abordagem estruturada baseada em técnicas de análise de dados e aprendizado de máquina, com ênfase na criação de modelos preditivos que

proporcionem uma visão aprofundada dos padrões e fatores de risco envolvidos no diagnóstico de câncer de mama.

As ferramentas utilizadas para a implementação e análise dos dados incluem a linguagem de programação Python e bibliotecas como Pandas e Numpy para manipulação de dados, Seaborn e Matplotlib para visualizações gráficas, e Scikit-learn para a criação e avaliação do modelo de regressão logística.

Inicialmente, a análise começou com uma exploração dos dados (EDA - Exploratory Data Analysis), buscando entender o comportamento das variáveis, suas distribuições e relações. Para isso, foram utilizados gráficos como gráficos de barras e gráficos de pizza, assim como a criação de uma matriz de correlação visualizada por meio de um *heatmap*. Essa etapa foi fundamental para identificar a relação entre as variáveis e a variável alvo, permitindo destacar as variáveis com maior potencial de influência no diagnóstico, como *radius_mean*, *perimeter_mean* e *concavity_mean*.

Para a construção do modelo preditivo, será aplicada a regressão logística, uma técnica amplamente utilizada em problemas de classificação binária, especialmente na área médica, por sua capacidade de prever a probabilidade de um evento ocorrer (neste caso, a malignidade do tumor). Essa técnica foi escolhida devido à sua simplicidade interpretativa e sua robustez na modelagem de relações entre múltiplas variáveis explicativas e uma variável de resposta binária.

Por fim, a investigação culminará com a apresentação dos insights obtidos, que podem contribuir para uma melhor compreensão dos fatores de risco associados ao câncer de mama e apoiar os profissionais de saúde em decisões clínicas mais informadas.

9. STORYTELLING

Nesse Capítulo, será abordado alguns tópicos importantes do trabalho como metadados AED e a resolução final, juntamente com uma breve descrição de como será realizada a apresentação.

- Grupo: Dunder Mifflin
- Integrantes: Gabriel Chaves Gonçalves
- Nome do Projeto: *UW Health University Hospital*

- Área de estudo: Saúde (oncologia mamária)
- Empresa: *UW Health University Hospital*
- Banco de dados: *Breast Cancer Wisconsin Dataset*
- Descrição do Problema: O objetivo é prever a malignidade do tumor, classificando-o como benigno ou maligno.
- Algoritmo Aplicado: Regressão Logística
- Resultados Esperados: Modelo preditivo robusto e confiável.

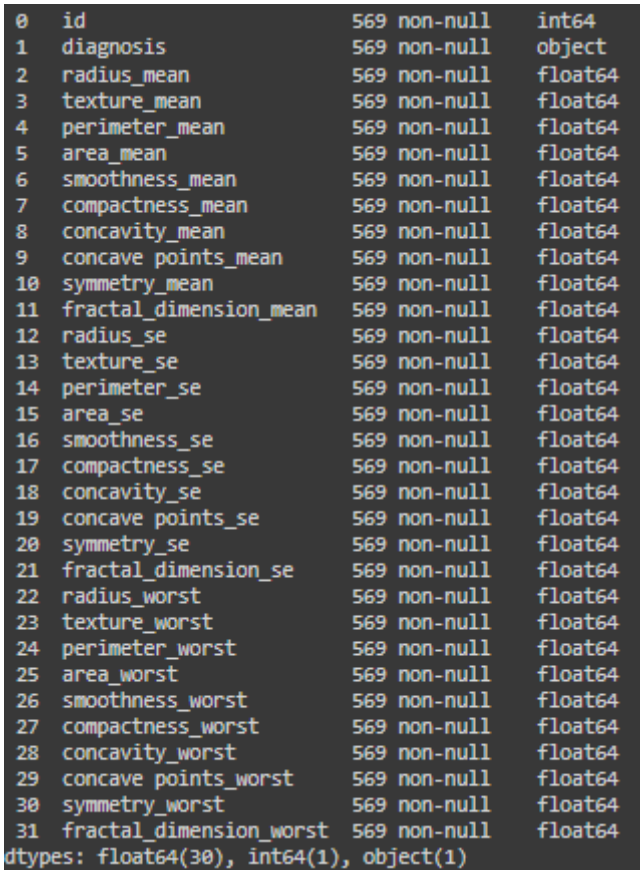
9.1. METADADOS

O *Breast Cancer Wisconsin Dataset* contém informações sobre tumores de mama, especificamente projetado para classificar os tumores em duas categorias: benignos e malignos. A proposta do *Breast Cancer Wisconsin Dataset* é fornecer um recurso que permita a pesquisa e o desenvolvimento de algoritmos de aprendizado de máquina para classificar tumores de mama com base em características quantitativas. O objetivo é utilizar essas informações para construir modelos que possam prever se um tumor é benigno ou maligno, facilitando o diagnóstico precoce e a decisão sobre o tratamento adequado.

O *dataset* aborda o fenômeno da detecção e classificação de tumores de mama, um desafio crucial na oncologia. Os problemas registrados incluem a necessidade de distinguir entre tumores benignos e malignos com alta precisão. A presença de características específicas do tumor, como tamanho e textura, é essencial para a classificação correta. O *dataset* também pode ajudar a identificar padrões e correlações que são úteis para a pesquisa clínica e o desenvolvimento de novas abordagens de tratamento.

O banco de dados está no formato csv e os dados são públicos disponíveis em diversos repositórios, como o Kaggle. Não contém dados sensíveis, pois foi identificado e não inclui informações pessoais identificáveis (PII) sobre os pacientes. Possui 569 registros e 32 campos. Dentre esses campos, 30 são do tipo float, 1 é do tipo int e 1 é do tipo object, permitindo uma análise qualitativa e quantitativa. Além disso, o dataset está isento de dados nulos ou faltantes, conforme ilustrado na Figura 1 abaixo.

Figura 1 – Característica geral dos dados



```
0   id                      569 non-null    int64
1   diagnosis               569 non-null    object
2   radius_mean             569 non-null    float64
3   texture_mean            569 non-null    float64
4   perimeter_mean          569 non-null    float64
5   area_mean               569 non-null    float64
6   smoothness_mean         569 non-null    float64
7   compactness_mean        569 non-null    float64
8   concavity_mean          569 non-null    float64
9   concave points_mean     569 non-null    float64
10  symmetry_mean           569 non-null    float64
11  fractal_dimension_mean  569 non-null    float64
12  radius_se               569 non-null    float64
13  texture_se              569 non-null    float64
14  perimeter_se            569 non-null    float64
15  area_se                 569 non-null    float64
16  smoothness_se           569 non-null    float64
17  compactness_se          569 non-null    float64
18  concavity_se            569 non-null    float64
19  concave points_se       569 non-null    float64
20  symmetry_se             569 non-null    float64
21  fractal_dimension_se    569 non-null    float64
22  radius_worst            569 non-null    float64
23  texture_worst           569 non-null    float64
24  perimeter_worst         569 non-null    float64
25  area_worst              569 non-null    float64
26  smoothness_worst        569 non-null    float64
27  compactness_worst       569 non-null    float64
28  concavity_worst         569 non-null    float64
29  concave points_worst    569 non-null    float64
30  symmetry_worst          569 non-null    float64
31  fractal_dimension_worst 569 non-null    float64
dtypes: float64(30), int64(1), object(1)
```

O conjunto de dados inclui as seguintes características:

- ID do Paciente: Identificador único para cada paciente.
- Classificação: Classe do tumor, que pode ser 'B' (benigno) ou 'M' (maligno).
- Características dos Tumores: Medições detalhadas das características dos tumores, incluindo:
- Raio (mean, se, worst): Tamanho do tumor.
- Textura (mean, se, worst): Uniformidade da textura do tumor.
- Perímetro (mean, se, worst): Comprimento ao redor do tumor.
- Área (mean, se, worst): Área total ocupada pelo tumor.
- Suavidade (mean, se, worst): Medida da suavidade da superfície do tumor.
- Compactação (mean, se, worst): Densidade do tumor.
- Concavidade (mean, se, worst): Grau de indentação na superfície do tumor.

- Pontos de concavidade (mean, se, worst): Número de pontos onde o tumor é mais côncavo.
- Simetria (mean, se, worst): Grau de simetria do tumor.
- Fractal Dimension (mean, se, worst): Complexidade da forma do tumor.

9.2. ANÁLISE EXPLORATÓRIA DOS DADOS

A AED nos permitiu explorar e visualizar dados de maneira a identificar tendências, anomalias e relações que podem não ser imediatamente evidentes. Através de técnicas de visualização, como gráficos de dispersão e matrizes de correlação, buscamos entender a distribuição das variáveis, a presença de outliers e a interdependência entre as características.

Para entender um pouco sobre as características dos dados que estava sendo trabalhada, iniciou-se entendendo a base de dados, que contém 569 registros e 32 campos. Dentre esses campos, 30 são do tipo *float*, 1 é do tipo int e 1 é do tipo *object*, permitindo uma análise qualitativa e quantitativa.

Depois disso procurou-se entender um pouco mais sobre as medidas descritivas dos dados como as tendências centrais como média e mediana, dispersão e limites de cada classe. As medidas descritivas trabalhadas foram a quantidade de valores, a média, desvio padrão, valor mínio, os quartis e valor máximo, conforme pode-se observar na figura 2.

Figura 2 – Exemplo de Características descritivas dos dados

	id	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean
count	5.690000e+02	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000
mean	3.037183e+07	14.127292	19.289649	91.969033	654.889104	0.096360	0.104341	0.088799	0.048919	0.181162
std	1.250206e+08	3.524049	4.301036	24.298981	351.914129	0.014064	0.052813	0.079720	0.038803	0.027414
min	8.670000e+03	6.981000	9.710000	43.790000	143.500000	0.052630	0.019380	0.000000	0.000000	0.106000
25%	8.692180e+05	11.700000	16.170000	75.170000	420.300000	0.086370	0.064920	0.029560	0.020310	0.161900
50%	9.060240e+05	13.370000	18.840000	86.240000	551.100000	0.095870	0.092630	0.061540	0.033500	0.179200
75%	8.813129e+06	15.780000	21.800000	104.100000	782.700000	0.105300	0.130400	0.130700	0.074000	0.195700
max	9.113205e+08	28.110000	39.280000	188.500000	2501.000000	0.163400	0.345400	0.426800	0.201200	0.304000

Como observa-se na figura 2, para a coluna ID, não é necessária a interpretação descritiva pois ele se comporta como uma chave primária de cada registro, sendo um valor único para cada linha.

Após a identificação dos valores descritivos para todos os campos, a fim de se entender melhor o *target*, foi realizado o cálculo da proporção entre tumores malignos e benignos e observou-se que essa proporcionalidade é de 62,74% para os tumores benignos e 37,26% para os tumores malignos e para elucidar de uma forma mais visual, foi construído um gráfico de barras e um gráfico de pizza mostrando essa proporção, conforme mostra as figuras 3 e 4.

Figura 3 - Gráfico de barras para proporção das classes dos Tumores

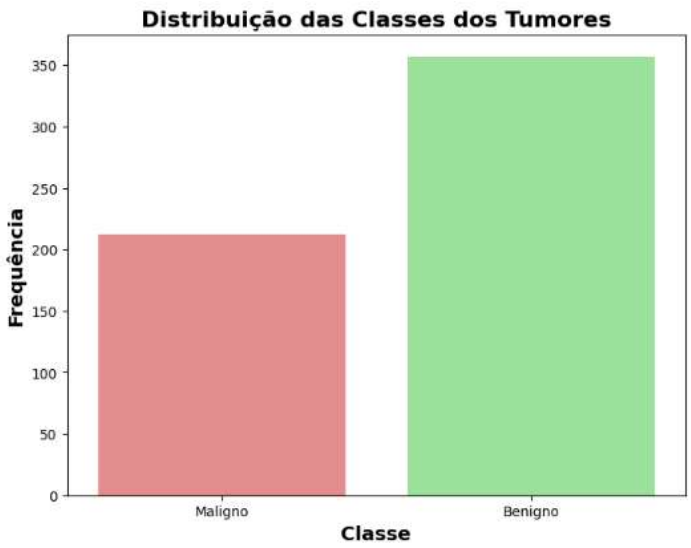
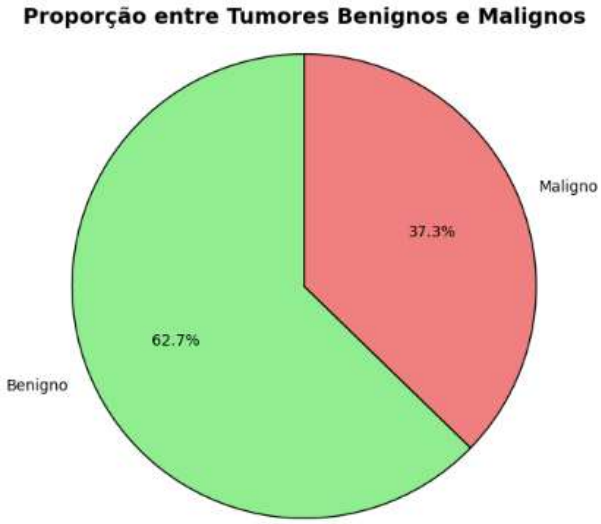
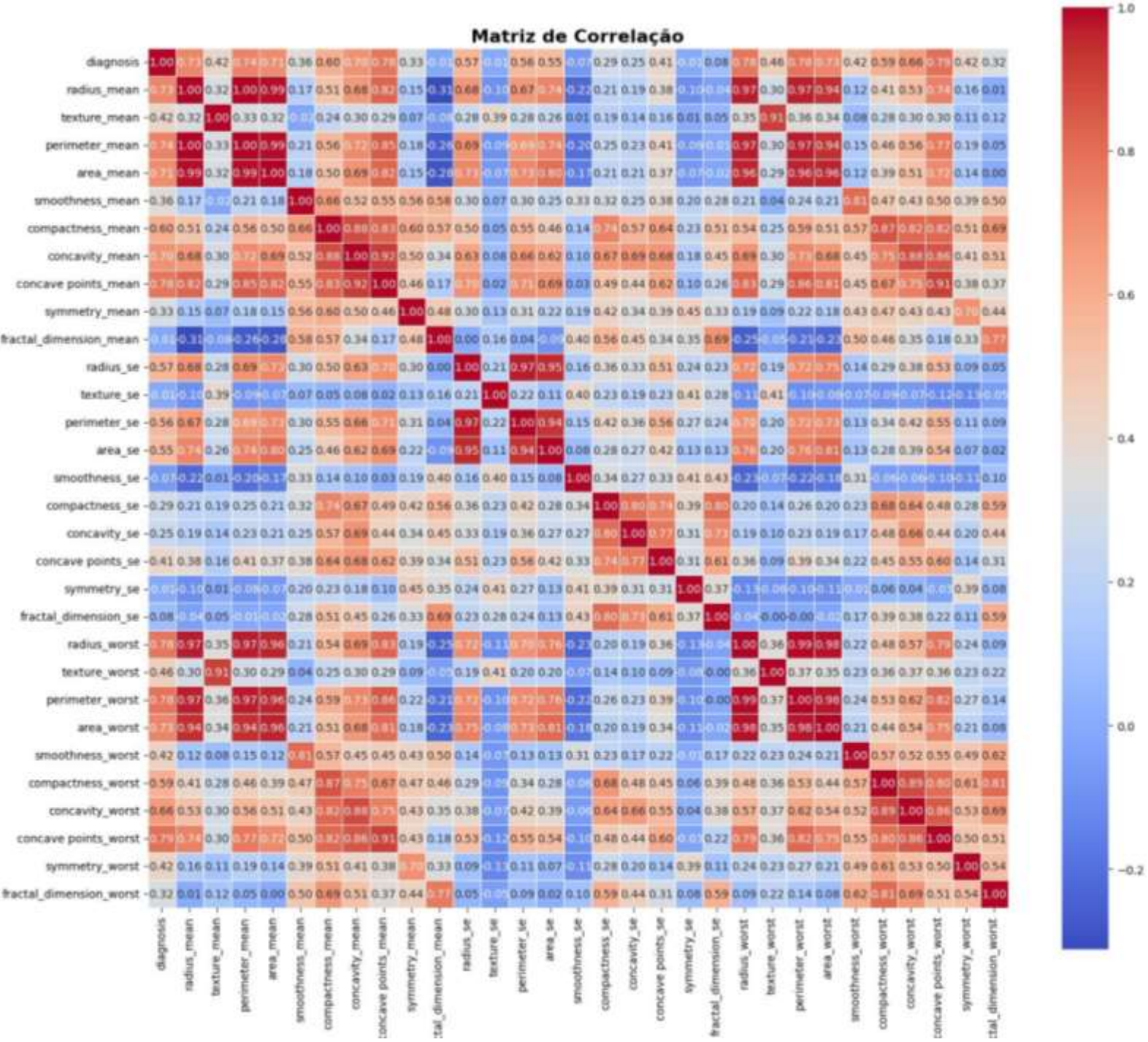


Figura 4 - Gráfico de pizza para proporção das classes dos Tumores



Uma vez que o comportamento da variável de interesse foi descoberto, entendeu-se a maioria das pacientes não possuíam câncer de mama. Então, foi realizado a análise de correlação entre as variáveis e pra isso foi desenvolvida uma mátria de correlação com um mapa de calor, conforme é possível observar na figura 5.

Figura 5 - Matriz de correlação entre as variáveis

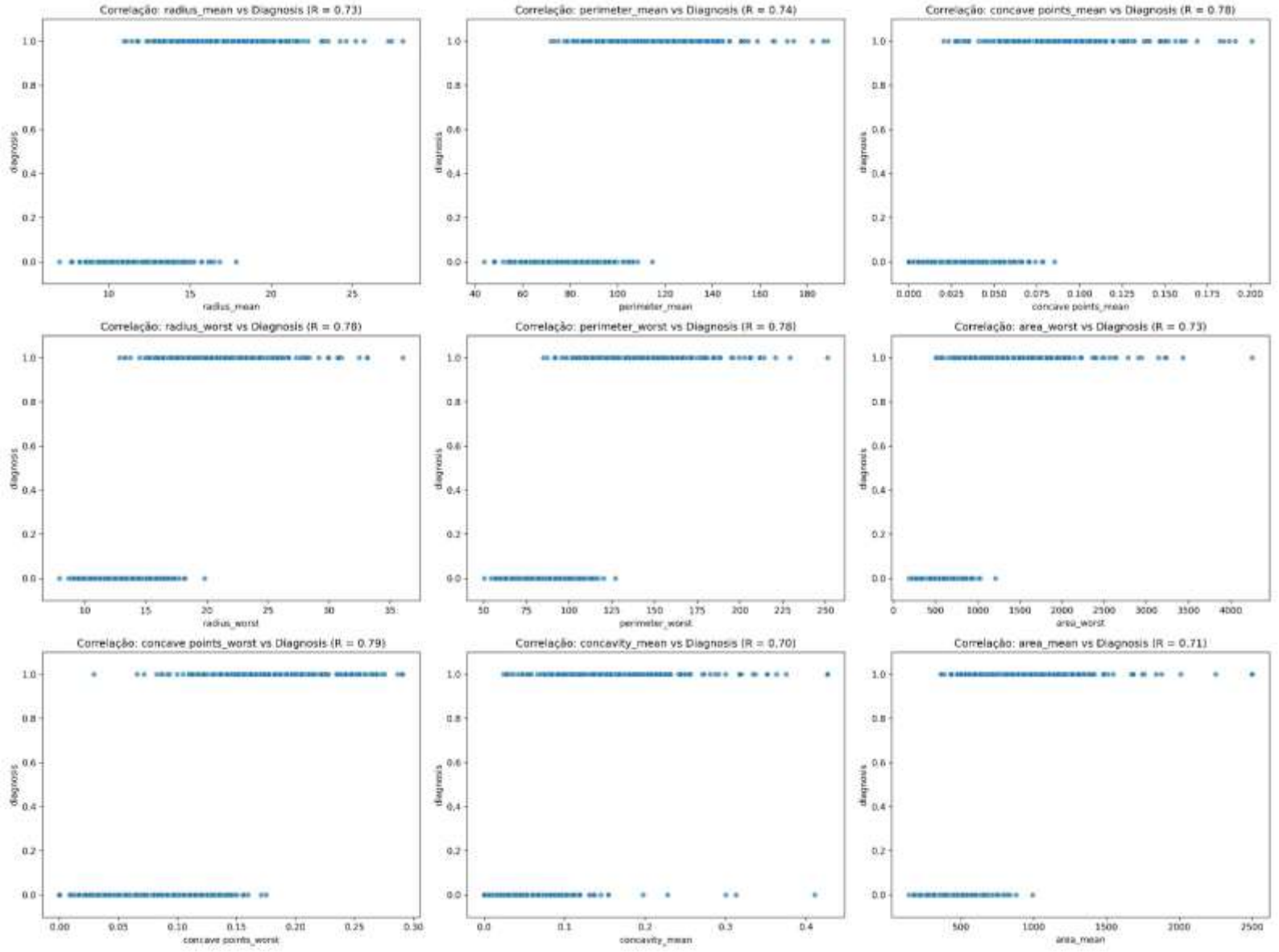


A matriz de correlação permitiu identificar as relações entre as variáveis do conjunto de dados, incluindo a variável alvo. Observa-se que algumas variáveis apresentaram uma forte correlação com a variável de interesse, com coeficientes de correlação de Pearson (R) variando entre 0,61 e 0,80. Esses valores indicam uma relação linear significativa, destacando-se como

características importantes para a análise, pode-se destacar as variáveis raio médio, perímetro médio, média do número de pontos côncavos, maior raio, maior perímetro, maior área e o maior número de pontos côncavos.

Para ilustrar o comportamento isolado entre as características e a variável alvo, foi construído um gráfico de dispersão e na figura 6, é possível observar como se dá essa correlação. De forma geral quanto maior os valores para essas características, maior será a probabilidade de o tumor ser maligno, que está representado pelo valor (1), no eixo y (diagnosis).

Figura 6 - Gráfico dispersão características vs variável alvo



Dada a alta correlação foi iniciado um estudo sobre a área de domínio, buscando entender o motivo dessa forte correlação.

- **Raio médio (radius mean)**

Tumores malignos tendem a ser maiores em tamanho, o que explica o fato de o raio médio ser mais elevado. Um raio maior indica um crescimento descontrolado das células.

- **Perímetro médio (perimeter mean)**

O perímetro mede o contorno do tumor. Tumores malignos possuem contornos mais irregulares e extensos, refletindo um maior perímetro médio.

- **Média do número de pontos côncavos (concave points mean)**

Tumores malignos têm bordas mais irregulares e deformadas, apresentando mais pontos côncavos. Isso é um sinal de crescimento anormal das células tumorais.

- **Concavidade média (concavity mean)**

Tumores malignos tendem a apresentar bordas mais irregulares, com formas menos definidas e mais áreas de concavidade, o que pode indicar um comportamento invasivo, onde as células cancerígenas estão se espalhando para os tecidos ao redor de maneira desordenada.

- **Maior raio (radius worst)**

O maior raio reflete o ponto de maior extensão do tumor. Tumores malignos geralmente atingem tamanhos maiores, o que torna essa medida importante para identificar malignidade.

- **Maior perímetro (perimeter worst)**

Como no perímetro médio, o maior perímetro também reflete o grau de irregularidade do tumor, sendo mais comum em tumores malignos devido ao seu crescimento desordenado.

- **Maior área (area worst)**

Tumores malignos geralmente ocupam uma área maior, pois crescem rapidamente. A maior área reflete a extensão desse crescimento.

- **Maior número de pontos côncavos (concave points worst)**

A presença de mais pontos côncavos nos tumores indica que as bordas são mais irregulares, características regularmente associadas ao câncer.

- **Área Média (área mean)**

Interpretação Clínica: A área de um tumor é uma medida do seu tamanho total. Tumores malignos, devido à sua capacidade de crescimento rápido e

descontrolado, tendem a ser maiores em comparação com tumores benignos.

9.3. ALGORITMO REGRESSÃO LOGÍSTICA

Para a etapa do uso de um modelo de aprendizado de máquina para a solução do problema, foi utilizado o algoritmo de Regressão Logística, pois a partir dele é possível observar não somente a classificação predita, mas também a probabilidade de ser a classificação predita.

A regressão logística é um método estatístico amplamente utilizado para modelar a relação entre um conjunto de variáveis independentes e uma variável dependente categórica, geralmente binária (com dois possíveis resultados, como "sucesso" e "falha" ou "0" e "1").

Diferentemente da regressão linear, que utiliza uma linha reta para modelar os dados, a regressão logística usa uma função logística ou sigmoide para produzir uma curva em "S", restrita entre 0 e 1. Essa característica permite que a regressão logística estime a probabilidade de um evento específico ocorrer.

A função da regressão logística, posso ser observada através de equação 1 e é expressa como:

Equação 1 - Função da Regressão Logística

$$P(Y=1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+...+\beta_n X_n)}}$$

Onde:

- P(Y=1|X): representa a probabilidade do evento de interesse ocorrer.
- Função Sigmoide $\frac{1}{1+e^{-z}}$**: converte a saída linear em uma probabilidade entre 0 e 1, criando a curva logística em "S". O valor de z é o somatório linear dos parâmetros multiplicados pelas variáveis independentes.
- e: o número de Euler (aproximadamente 2.718), que é a base dos logaritmos naturais.

- $\beta 0$: o intercepto do modelo, representando o valor da função quando todas as variáveis independentes são zero. Ele ajusta a posição da curva logística ao longo do eixo Y.
- β1,β2,…,βn: os coeficientes para cada variável independente.
- $X1, X2, \ldots, Xn$: as variáveis independentes ou preditoras.

A função logística transforma a combinação linear das variáveis em uma probabilidade, o que facilita a classificação. É um método bastante eficaz para problemas de classificação binária e é utilizado em áreas como medicina, finanças e ciência de dados para prever resultados binários com alta precisão.

9.4. RESOLUÇÃO FINAL

Para realizar a aplicação do algoritmo foram selecionadas as características com correlação linear superior a 0.69. Logo, as características escolhidas foram:

- Raio médio
- Perímetro médio
- Média do número de pontos côncavos
- Maior raio
- Maior perímetro
- Maior área
- Maior número de pontos côncavos.
- Área média
- Concavidade média

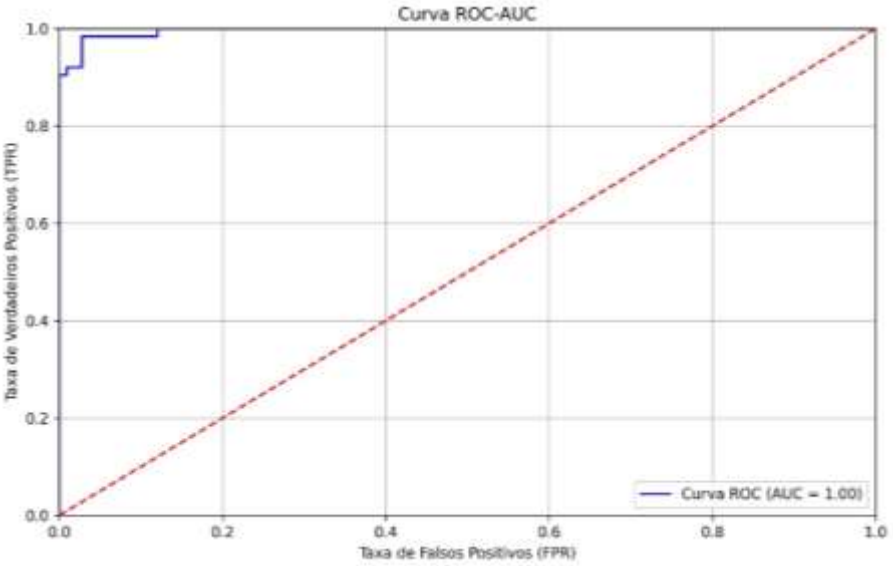
Como explicado anteriormente no Capítulo 11, o algoritmo utilizado foi a Regressão Logística. Neste projeto, a aplicação foi realizada com sucesso, resultando em um modelo com alta precisão para classificar tumores mamários em benignos e malignos.

As métricas escolhidas para avaliar o modelo foram Acurácia e a Curva ROC-AUC. A Acurácia foi utilizada por retornar à proporção de previsões corretas, sendo adequada uma vez que os dados, conforme mostrado na Figura 3, não estão fortemente desbalanceados.

A Curva ROC-AUC foi incluída por sua capacidade de avaliar a separação entre as duas classes de forma abrangente. Esta métrica é ideal para classificações binárias, especialmente porque é independente de limiares específicos, fornecendo um único valor representativo e permitindo análise visual e comparativa da performance do modelo.

Com uma acurácia de 96,49% e AUC-ROC de 99,6%, o modelo demonstrou uma capacidade robusta de distinguir entre classes, o que é essencial em diagnósticos médicos. Na figura 7, é possível observar o comportamento da curva ROC-AUC.

Figura 7 - Curva ROC-AUC



A linha azul representa a curva ROC do seu modelo, e a linha pontilhada vermelha mostra a linha de referência (que seria uma classificação aleatória). A área sob a curva (AUC) é próxima a 1, indicando um desempenho ideal. Esse desempenho de classificação permite decisões mais eficazes, impactando diretamente na qualidade do tratamento.

A acurácia média de 94,72% na validação cruzada, acompanhada de um baixo desvio padrão de 0,0177, reforça a confiabilidade do modelo em diferentes amostras, indicando baixo risco de *overfitting*. Essa estabilidade do modelo é crucial para que ele possa ser utilizado em novos conjuntos de dados com variações mínimas, aumentando a confiança no uso clínico. Pois afirmações falsas podem ser custosas no âmbito da saúde.

Ainda a fim de se testar o modelo, foi criado um sistema para imputar valores aleatórios para as 9 características mais influentes, dessa maneira o usuário final poderia realizar novos testes e a partir do modelo construído será gerado a probabilidade de o tumor ser maligno ou benigno.

A opção por uma abordagem probabilística mostrou-se relevante no contexto clínico, pois fornece uma estimativa contínua do risco de malignidade, que auxilia os médicos na priorização de exames e tratamentos. Esse recurso permite decisões mais personalizadas e alocações de recursos mais direcionadas, essencial em oncologia onde tempo e precisão são vitais.

Além disso, o modelo oferece benefícios amplos para a área da saúde, contribuindo para uma medicina mais personalizada e focada na prevenção. A avaliação probabilística permite monitoramento preventivo e detecção precoce de condições graves, aumentando a eficiência do sistema de saúde como um todo, reduzindo custos e melhorando a gestão dos cuidados.

9.5. APRESENTAÇÃO

A apresentação será dividida em cinco partes. Primeiro, abordaremos e apresentaremos a empresa e o tipo de problema tratado. Em seguida, discutiremos o banco de dados selecionado e a análise exploratória realizada. Por fim, apresentaremos o modelo criado, destacando como ele responde aos objetivos do trabalho e as vantagens de sua aplicação na medicina oncológica.

10. REFERÊNCIAS

CONSOLI, S.; REFORGIATO RECUPERO, D.; PETKOVIĆ, M.;(Eds.). Data Science for Healthcare: **Methodologies and Applications**. Cham: Springer, 2019.

DAVENPORT, T. KALAKOTA, R. (2019) The Potential for Artificial Intelligence in Healthcare. **Future Healthcare Journal**, 6, 94-98.

PEARSON, K. **Mathematical contributions to the theory of evolution. III. Regression, heredity, and panmixia**. Philosophical Transactions of the Royal Society of London, London, v. 187, p. 253-318, 1896.

HOSMER, David W.; LEMESHOW, Stanley; STURDIVANT, Rodney X. **Applied Logistic Regression**. 3. ed. Hoboken, NJ: John Wiley & Sons, 2013.

MENARD, Scott. **Logistic Regression: From Introductory to Advanced Concepts and Applications**. Thousand Oaks, CA: SAGE Publications, 2010.

11. LINK GITHUB

https://github.com/GabrielGoncalvesG/Projeto-Aplicado-I.git

12. APRESENTAÇÃO YOUTUBE

https://youtu.be/IJvE6ek1RSc