Análise Estatística acerca do Câncer de Mama

Bernardo Vargas da Motta Alves

Nesse projeto, o dataset de dados sobre câncer de mama extraídos da Universidade de Wisconsin, EUA, será utilizado como base para análise estatística acerca de diferentes informações obtidas através de diversas imagens de pacientes com tumores, acompanhadas de categorização de tais tumores como beningnos ou malignos. O objetivo, além de entender a correlação das feaures com o eventual diagnóstico, é conseguir fazer predições de novos dados, ando uma resposta sobre a natureza do tumor tendo os valores das diversas medições como dados.

1 Introdução

Segundo dados da Organização Mundial da Saúde, no ano de 2020 mais de duas milhões de mulhres foram diagnosticadas com câncer de mama e cerca de 685 mil mortes foram registradas mundailmente. Além disso, nesse mesmo ano, existiam 7.8 milhões de mulheres vivas com algum diagnóstico de câncer de mama nos cinco anteriores. Em comparação com qualquer outro tipo de cãncer, o de mama é o que possui a pior "pontuação"no índice Disability-Adjusted Life Years (DALYs), que mede tanto os anos perdidos por conta de uma morte prematura quanto os anos vividos com incapacidades.

Apesar de as taxas de sobrevivência do câncer de mama terem crescido a partir da década de 80, os dados apresentados deixam clara a importância de um combate e a relevância e impacto que essa doença causa tanto em suas vítimas quanto nas famílias, motivando o presente projeto, que busca encontrar relações e padrões entre medições médicas e diagnóstico de tumores como malignos ou benignos.

Através de imagens realizadas em diversas mulheres com tumores, a Universidade de Wisconsin, EUA, disponibilizou dados que mediram dez métricas diferentes: radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry e fractal dimension. Três "espécies"de medições diferentes foram feitas para todas essas métricas, a média de cada uma, o erro padrão e a média dos três "piores"resultados, totalizando 30 métricas na base de dados que será usada para construção de modelos e realização de predições em cima de tais modelos.

Bernardo Vargas da Motta Alves: bernardovma@gmail.com

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430

Figura 1: Tabela com os Dados.

Com a figura acima, podemos perceber como foi estruturado o database, com um id como chave de identificação das pacientes e as medições vindo posteriormente, como o nome da medida seguido de mean quando se trata da média, como exposto no exemplo, e se e worst, quando se refere ao erro padrão e ao pior caso, respectivamente.

Devido a relevância de um diagnóstico como maligno ou benigno para decisões de vida de mulheres que passam por essa situação, alguns pontos são de importante consideração: por mais que um diagnóstico previsto maligno quando o diagnóstico correto é benigno seja indesejado pelas grandes decisões de vida que devem ser tomadas, é provável que exames posteriores comprovatórios sejam realizados que permitam um novo diagnóstico. Contudo, afirmar que uma mulher possui um tumor benigno quando na verdade este é maligno pode acarretar em um atraso no tratamento que piore a condição de saúde da paciente. Por esse motivo, por mais que a análise acerca dos casos de falso positivo seja de extrema importância para a avaliação dos modelos, os casos de falso negativo teram um enfoque maior nessas avaliações.

Dois caminhos principais serão tomados: i. escolha aleatória de quais e como as váriaveis serão introduzidas no modelo; ii. contrução de três modelos que utilizem todas as variáveis relacionadas à média, ao erro padrão e aos piores casos, respectivamente. Com a primeira abordagem, será possível fazer comparação entre diferente modelos e ver como a escolha das variáveis impacta nas predições e nas medidas de eficácias escolhidas. Já na segunda abrodagem, é possível perceber como os três diferentes grupos de dados se comportam em relação ao diagnóstico final como maligno ou benigno.

No fim do projeto, com a escolha do modelo, ajuste e comparação de diversos modelos, exposição de resultados, é desejado que seja encontrado um modelo satisfatório, com boas respostas nas medições escolhidas, consequência de uma boa capacidade preditiva.

2 Metodologia

2.1 Modelos

A primeira grande decisão metodológica do projeto foi pela escolha de utilização de modelos de regressão logóstica. Os dados possuem variável dependente categórica binária (diagnóstico como maligno ou benigno), incetivando a utilização da regressão logística, gerando saídas que são probabilidades dos casos analisados corresponderem a uma ou outra classe dentro da variável dependente. Essa razão adicionada ao fato de podermos utilizá-la mesmo não existindo uma relação linear entre as variá-

veis independentes e as dependentes, a escolha da regressão logística foi feita como modelo a ser usado para os dados.

Como exposto, as esolhas das variáveis a serem escolhidas para integrar os modelos e como elas se relacionaram, será feita de maneira aleatória, com o enfoque sendo na comparação entre diferentes modelos que possibilite a identificação daqueles com melhores resultados. Seguem três modelos iniciais escolhidos:

- Modelo Simples Inicial (Modelo 1):
  - utilizando média do raio (*radius-mean*) + a média do perímetro (*perimeter-mean*).
- Modelos mais incrementados:
  - Modelo 2 - soma da média do raio(*radius-mean*) e do erro padrão do raio (*radius-se*) multiplicada pelo pior caso do raio (*radius-worst*) + a soma da média do perímetro (*perimeter-mean*) com o erro padrão do perímetro (*perimeter-se*) multiplicada pelo pior caso do perímetro (*perimeter-worst*) + média da concavidade (*concavity-mean*) somada ao erro padrão da concavidade (*concavity-se*) multiplicada ao pior caso da concavidade (*concavity-worst*)
  - Modelo 3 - soma da média da suavidade(*smoothness-mean*) e do erro padrão da suavidade (*smoothness-se*) multiplicada pelo pior caso da suavidade (*smoothness-worst*) + a soma da média dos pontos côncavos (*concave-points-mean*) com o erro padrão dos pontos côncavos (concave-points-se) multiplicada pelo pior caso dos pontos côncavos (concave-points-worst) + média da simetria (*symmetry-mean*) somada ao erro padrão da simetria (*symmetry-se*) multiplicada ao pior caso da simetria (*symmetry-worst*)

Para a outra parte das análises, serão ajustados modelos que utilizam as divisões em grandes grupos das variáveis dependentes:

- Modelo Mean - utilizando todas a variáveis independentes relacionadas à Média.
- Modelo SE - utilizando todas a variáveis independentes relacionadas ao Erro Padrão.
- Modelo Worst - utilizando todas a variáveis independentes relacionadas ao Pior Caso.

2.2 Ajuste

Como será feita a utilização da biblioteca '*statsmodel*', no python, para os ajustes dos modelos, é importante ressaltar que todos os modelos da biblioteca são estimados utilizando máxima verossimilhança e assumindo erros distribuídos identica e independentemente. Então podemos ver que será utilizada uma abordagem frequentista, estimados maximizando a função de verossimilhança, em contraste a uma análise bayesiana.

2.3 Avaliação

Uma importante parte na hora de comparar os modelos é escolher as medidas de avaliação que serão utilizadas para medir a eficácia/eficiência de cada um deles. Além de uma predição feita em cima dos dados que temos para depois realizarmos ocmparações com os resultados reais, medidas com enfoque na análise de falsos positivos e negativos foram utilizadas.

Como dito anteriormente, a saída de uma regressão logística é uma probabilidade, e de forma exterior ao ajuste do modelo, podemos definir quando a linha de divisão dentro dessas probabilidades que faz com que a classificação seja feita como Maligno ou Benigno, dando margem para escolhas de projeto acerca de priorização dos resultados. Atrelado a essas possíveis decisões, serão utilizadas técnicas como ROC-AUC e matiz de confusão para melhores análises.

O ROC (*Receiver Operating Characteristic*) é uma medida que visa mostrar capacidade de modelos em distinguir entre duas categorias (nesse caso maligno e benigno) atarvés das taxas de verdadeiro positivo e falso positivo em diferentes limiares de classificação. O AUC serve como tentativa de simplificar o ROC, criando um score, uma métrica, para essa medida, variando de 0 a 1, em que 1 é o valor desejado, pois significa que o modelo está classificando cem por cento das predições corretamente.

Já a matriz de confusão indica detalhes de todas as predições, indicando aqueles que foram classificadas como malignas corretamente e aquelas que deveriam ser classificadas como benignas e vice-versa. A partir desses dados, podemos definir e calcular alguns scores:

- Acurácia: taxa das classificações corretas do modelo em relação a todos os dados;
- Precisão: proporção de casos positivos corretamente classificados dentro do total de casos classificados como positivos;
- Recall/Sensibilidade: proporção de casos positivos corretamente classificados em relação ao total de casos cuja classificação esperada é positivo;
- Especificidade: proporção de casos negativos corretamente classificados em relação ao total de casos cuja classificação esperada é negativo;
- F1-score: medida que combina a sensibilidade e a precisão em um único valor.

Analisando as explicações dessas medições e remetendo a importância da diminuição dos falsos negativos, o recall e o f1-score se mostram adequadas para terem um destaque ainda maior, visto que o Recall representa exatamente 1 - a quantidade de falsos negativos sobre todos os negativos e o f1-score relaciona essa medida com a precisão, que visa identificar os positivos corretamente classificados dentro dos positivos que esperavam terem sido classificados. Assim, podemos ter uma boa análise dos modelos que atendem os requisitos propostos pelo projeto.

3 Reultados

3.1 Análise Exploratória dos Dados

Após a transformação dos dados em um dataframe, utilizando a biblioteca *pandas*, foi feita uma análise exploratória utilizando python, a fim de entender um pouco melhor os dados, ajudando na escolha dos modelos. O primeiro passo foi um plot de um gráfico de pizza mostrando a quantidade de diagnósticos por categoria. Apesar desse tipo de gráfico não ser recomendado para muitas situações, nesse caso dá uma fácil e rápida visualização que transmite bem a mensagem que pretende, podendo ser perceptível a maioria (aproximadamente dois terços) de tumores benignos.

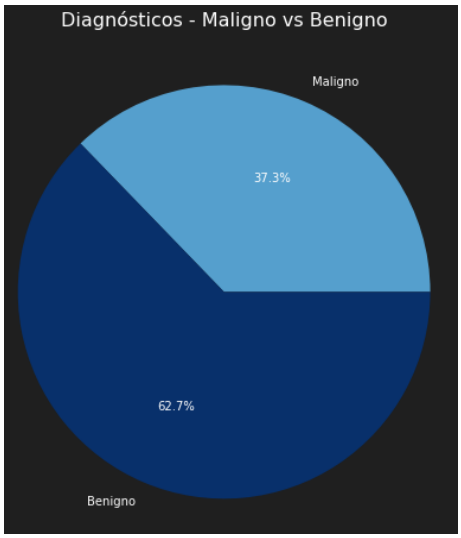

Figura 2: Porcentagem por Diagnóstico.

Após isso, foram plotados alguns gráficos mostrando relação entre variáveis independentes com a dependente, seguindo abaixo apenas um exemplo para mostrar uma dessas correlações, visto que são muitas métricas e seria impraticável mostrar muitas delas:

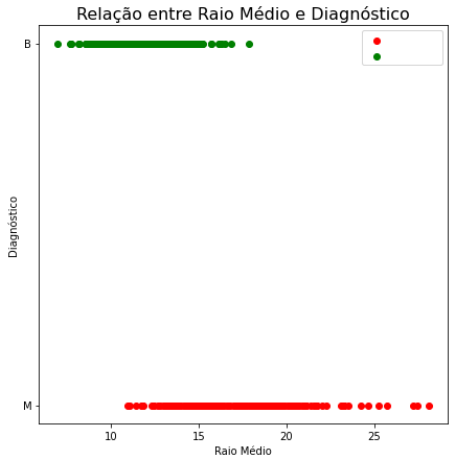

Figura 3: Correlação Raio Médio x Diagnóstico.

Completando com uma substituição de valores NaN por zero, que na verdade se mostrou irrelevante pois as únicas ocorrências eram de uma coluna que estava toda nula, e transformação dos valores do diagnóstico de maligno e benigno para 1 e 0, respectivamente, as dados estavam preparados para serem utilizados no ajuste de modelos, como será exposto a seguir.

Primeiro serão exibidos os resultados de todos os modelos ajustados para depois serem realizadas comentários e comparações.

3.2 Primeiro Modelo

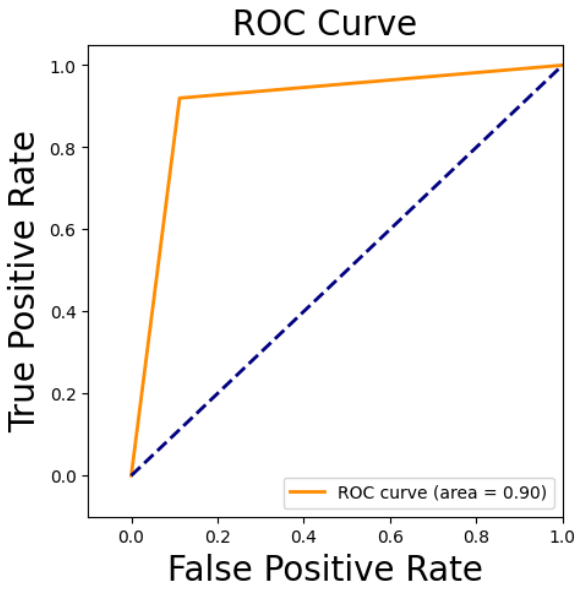

Figura 4: ROC-AUC Modelo 1.

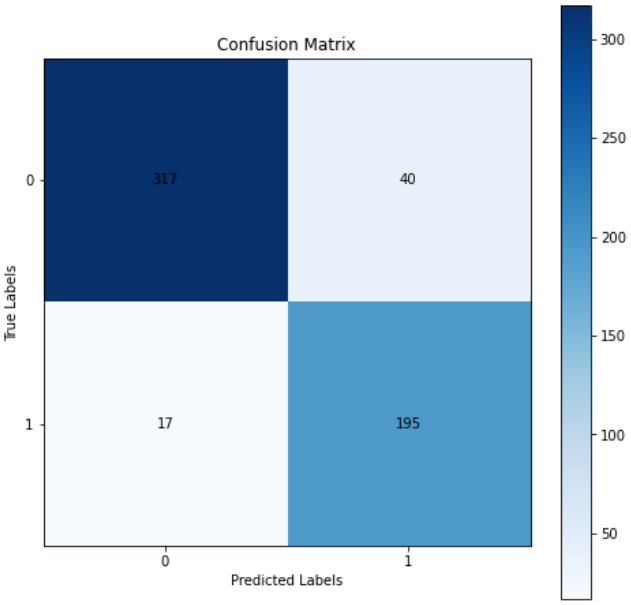

Figura 5: Matriz de Confusão Modelo 1.

3.3 Segundo Modelo

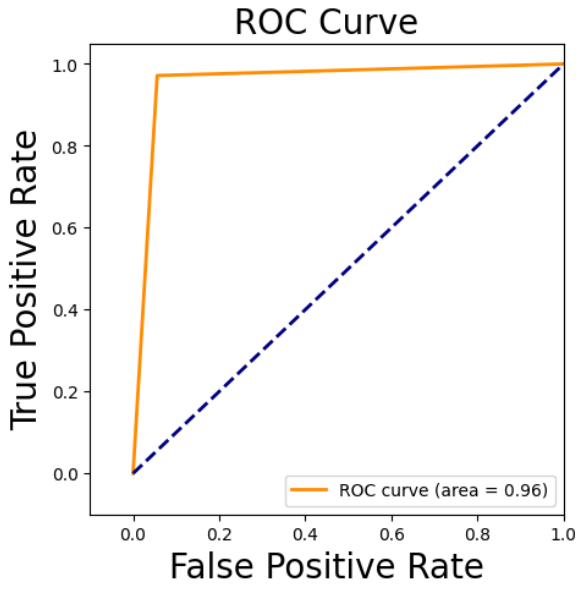

Figura 6: ROC-AUC Modelo 2.

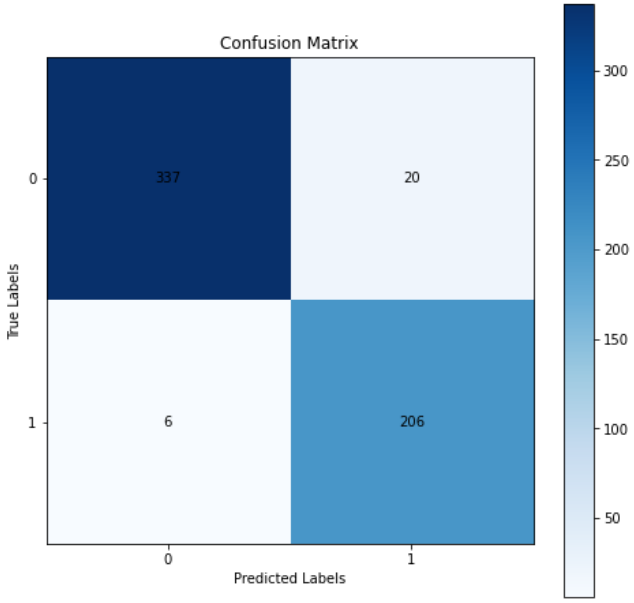

Figura 7: Matriz de Confusão Modelo 2.

3.4 Terceiro Modelo

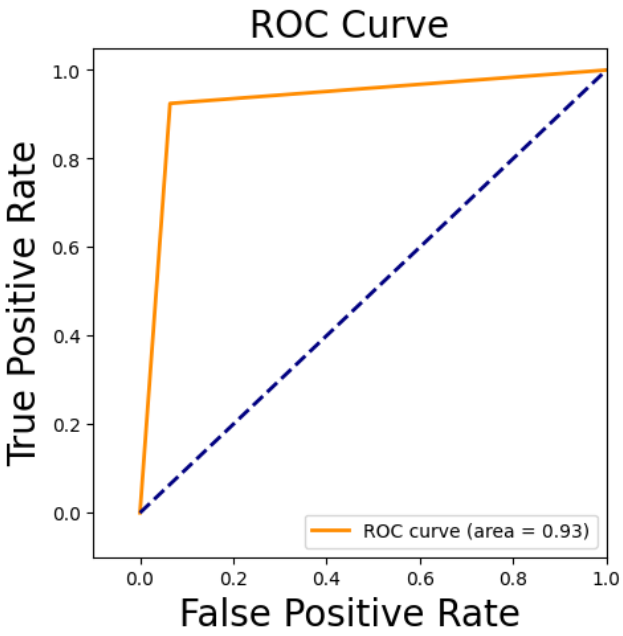

Figura 8: ROC-AUC Modelo 3.

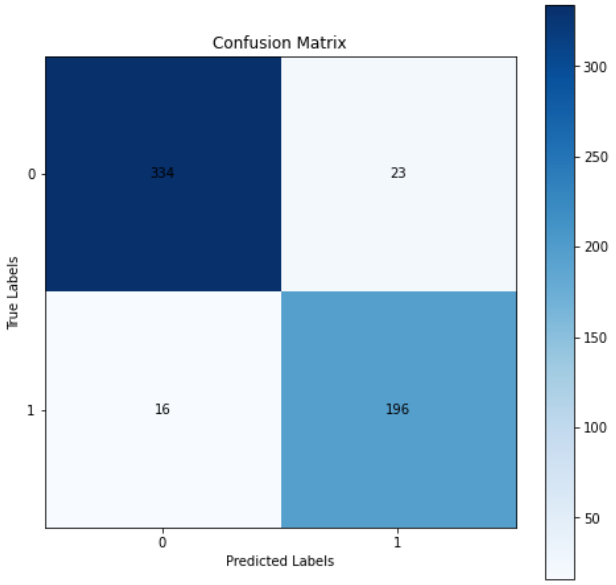

Figura 9: Matriz de Confusão Modelo 3.

3.5 Comparação dos Modelos Aleatórios

Inicialmente, é perceptível a qualidade dos modelos, até daquele mais simples com a simples soma de duas variáveis, com um bom score na curva e uma matriz satisfatória, indicando uma boa acurácia. Percebe-se também que isso se mantém para os dois próximos modelos, principalmente o segundo, que são um pouco mais complexos. A qualidade dos modelos pode ser visto na tabela a seguir:

	Modelo	AIC	Acurácia	Especificidade	Recall	Precisão	F1-Score
0	Modelo 1	256.993783	0.899824	0.887955	0.919811	0.829787	0.872483
1	Modelo 2	130.265186	0.954306	0.943978	0.971698	0.911504	0.940639
2	Modelo 3	197.682897	0.931459	0.935574	0.924528	0.894977	0.909513

Figura 10: Comparação entre Modelos Aleatórios.

Com todos os modelos com ótima acurácia, vemos um segundo modelo com pontuação excelente em todas as métricas, inclusive naquelas de destaque como o recall e o f1-score, sendo excelente na minimização dos falsos negativos que devem ser evitados. Um aspecto importante é que no momento de definir os valores de probabilidades que iriam se transformar em 0 (benigno) ou maligno (1), foi escolhido um valor de 0.3, ou seja, caso esse valor fosse maior que 0.3, era classificado como 1 e não como 0. Isso foi importante porque, mesmo que possa causar uma queda na acurácia ou aumento em falsos positivos, ajuda a reduzir os falsos negativos e o diagnóstico equivocado de alguém que possui um tumor maligno que deve já ser tratado.

3.6 Modelo Mean

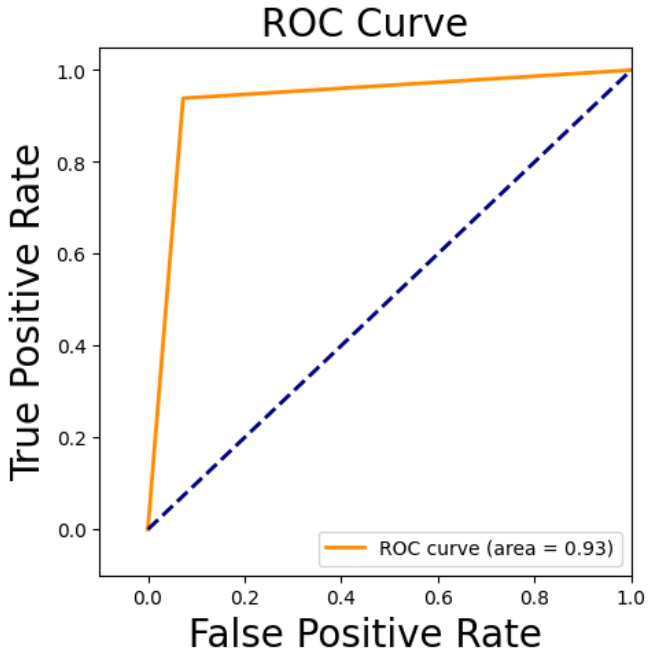

Figura 11: ROC-AUC Modelo Mean.

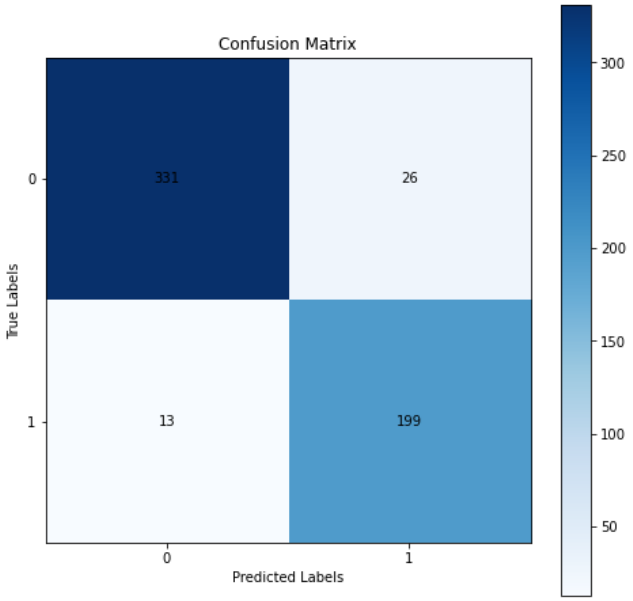

Figura 12: Matriz de Confusão Modelo Mean.

3.7 Modelo SE

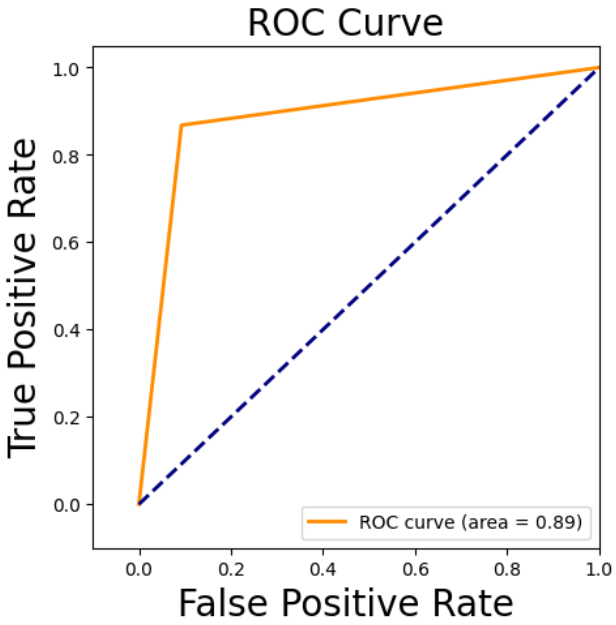

Figura 13: ROC-AUC Modelo SE.

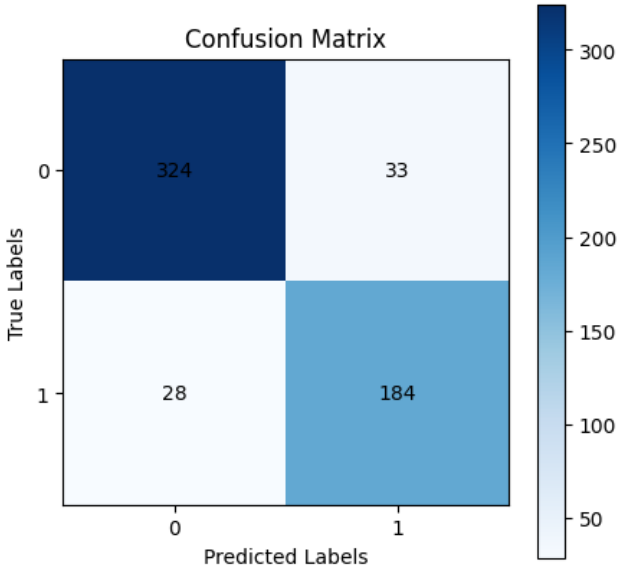

Figura 14: Matriz de Confusão Modelo SE.

3.8 Modelo Worst

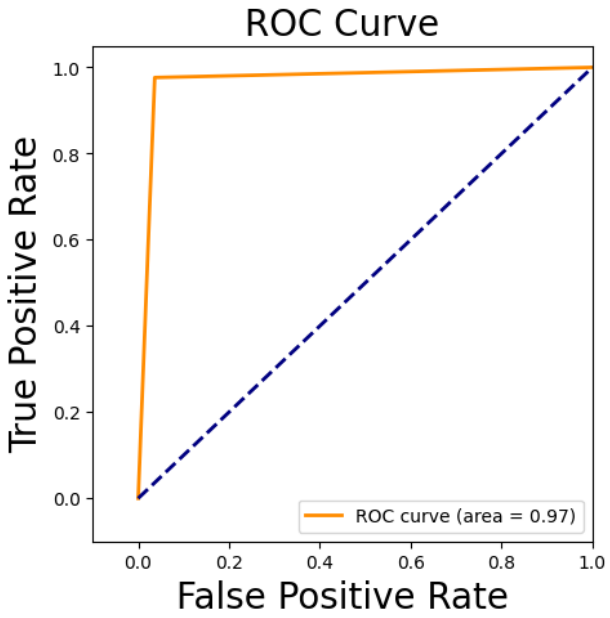

Figura 15: ROC-AUC Modelo Worst.

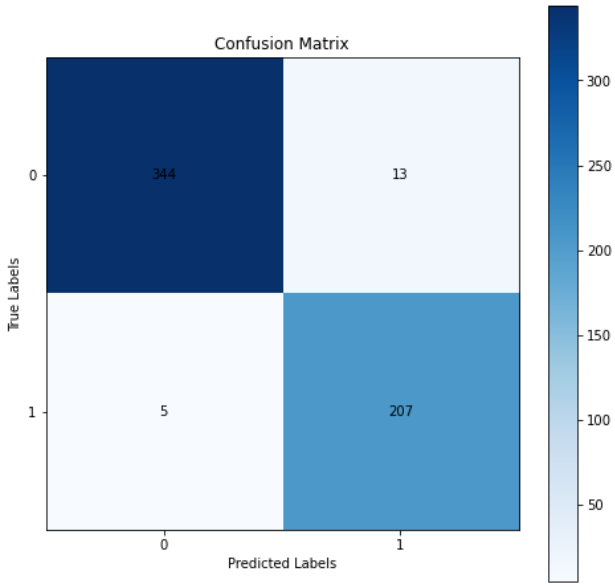

Figura 16: Matriz de Confusão Modelo Worst.

3.9 Comparação Modelos Mean/SE/Worst

Novamente com uma qualidade impressionante de todos os modelos, temos a tabela comparativa:

	Modelo	AIC	Acurácia	Especificidade	Recall	Precisão	F1-Score
0	Modelo Mean	168.130418	0.931459	0.927171	0.938679	0.884444	0.910755
1	Modelo SE	287.917537	0.892794	0.907563	0.867925	0.847926	0.857809
2	Modelo Worst	105.567018	0.968366	0.963585	0.976415	0.940909	0.958333

Figura 17: Comparação entre Modelos Mean/SE/Worst.

Foi mantida a estrutura das probabilidades e percebe-se um modelo ainda melhor que o Modelo 2, o Modelo Worst, com resultados levemente melhores de algo que já era muito bom. Vale destacar a aparição da métrica AIC nas tuas tabelas, que é mais uma medida comparativa entre modelos e serve justamente para comparar modelos do mesmo tipo, com o valor absoluto sendo de segunda importância em relação a comparação desse valor entre os diferentes modelos.

4 Conclusão

Com os resultados obtidos acima, podemos destacar a qualidade dos dados em relação às predições realizadas, com modelos simples gerando resultados surpreendentes. Nem apenas na acurácia, que é uma medida importante, mas não determinante em todos os casos, as métricas escolhidas como relevantes também foram bem representadas em todos os modelos, com os valores mais baixos chegando na faixa de 0.87 (num socre de 0 a 1). Visto que o esperado era o teste de vários modelos com uma grande variação de resultados, indo de bons a ruins, foi inesperada essa qualidade imediata.

Além disso, a segunda parte da anáise mostrou uma característica esperada, com as piores medições sendo mais determinantes e relacionadas com o diagnóstico. Apesar dos bons resultados das outras categorias, com a média sendo melhor do que o erro padrão, as piores medições formaram o melhor modelo entre todos aqueles ajustados, com acurácia acima de 0.95 e uma taxa de falsos negativos abaixo de 0.03 (recall 0.97), sendo super adequada a natureza dos dados e o foco do projeto.

Por mais que as bondades do ajuste escolhidas, as métricas e avaliações não trem sido as mais complexas possíveis, foi uma importante lição escolher o foco na redução de falsos negativos pois possibilitou encontrar as métricas ideias para identificar a qualidade dos modelos ajustados. A simples utilização da acurácia demonstra, às vezes, um certo despreparo para lidar com um ajuste de um modelo estatístico, visto que a natureza dos dados e da análise pedem por medidas mais específicas e especializadas para casos particulares.

A grande quantidade de váriaveis independentes mostra-se uma solução e um problema ao mesmo tempo. Por mais que libere a possibilidade de testar vários modelos, também acarreta em diversos ajustes diferentes de não srem realizados, fazendo com que alguns ótimos modelos possam ficar perdidos. Testar modelos de forma mais padronizada, e não aleatória, pode ser um próximo passo importante para extrair omáximo dos dados presentes. Além disso, mais dados seriam ideais, pois um pouco menos de 600 linhas são uma boa quantidade, mas com mais dados é possível ajustar ainda melhor os modelos escolhidos e melhorar ainda mais a análise.

Com um problema latente em mãos e dados de qualidade, foi possível fazer uma análise de acordo com essa qualidade, encontrando padrões e inferências interessan-

tes que ajudam a contar a história e que podem auxiliar no combate dessa doença que prejudica mulheres e seus familiares por todo o mundo.

Referências

[1] Aprofundamento de Métricas em "Métricas Precision/Recall/F1-Score"

[2] Modelos no Statsmodel em "Modelo de Regressão Logísica no Statsmodel"

[3] Regressão Logística em "Regression and Other Stories"