Udacity Machine Learning Engineer Nanodegree

Capstone Project:
Diagnósticos de câncer de mama através de algoritmos de machine learning.

Aluno: Jean Shinohara
21/abril/2018

Index

Introdução

Antes de entender o que é o câncer de mama em si, é preciso entender o que caracteriza um câncer.

O corpo humano é composto por diversos tecidos que, consequentemente, são compostos por diversas células. Um câncer é caracterizado por um crescimento acelerado e desordenado dessas células, independentemente da parte do corpo, causando um tumor. Quando não diagnosticado e tratado precocemente, esse tumor tende a crescer de forma muito agressiva, tornando-se, assim, maligno.

Tendo esse panorama, pode-se dizer então que o câncer de mama, como o próprio nome já diz, afeta as células da mama, glândulas constituídas pelas seguintes estruturas:

- Estruturas produtoras de leite (lóbulos);
- Ductos (pequenos canais que ligam os lóbulos ao mamilo);
- Gordura;
- Tecido conjuntivo (tecido que liga, nutre, protege e sustenta outros tecidos);
- Vasos sanguíneos;
- Vasos linfáticos.

Normalmente, esse tipo de câncer tem início nos lóbulos ou nos ductos. Porém, ele pode ser iniciado também em outras estruturas das mamas, como os tecidos estromais, tecido que inclui as gorduras e o tecido conjuntivo da glândula.

Segundo o INCA (Instituto Nacional de Câncer), o câncer de mama é o tipo de câncer que mais acomete as mulheres e responde por 25% dos casos novos da doença no Brasil. Todavia, a doença pode acometer os homens também, porém de forma bem mais rara; apenas 1% dos homens brasileiros possuem o tumor.

Dados do INCA afirmam que, no ano de 2013, o câncer de mama levou a óbito 14.388 pessoas. Delas, 181 eram homens e o restante, isto é, 14.206, eram mulheres. Ainda segundo o Instituto, será computado em 2016 cerca de 58 mil novos casos da doença.

Não existe uma única causa para o câncer de mama. O que acontece, na verdade, é a influência de um conjunto de fatores que pode desencadear o início da doença.
Um fator de risco é algo que afeta sua chance de adquirir uma doença como o câncer. Diferentes tipos de câncer apresentam diferentes fatores de risco. Alguns como fumar, por exemplo, podem ser controlados; no entanto outros não, por exemplo, idade e histórico familiar.

Embora os fatores de risco possam influenciar o desenvolvimento do câncer, a maioria não causa diretamente a doença. Algumas pessoas com vários fatores de risco nunca irão desenvolver um câncer, enquanto outros, sem fatores de risco conhecidos poderão fazê-lo.

Ter um fator de risco ou mesmo vários, não significa que você vai ter a doença. Muitas pessoas com a enfermidade podem não estar sujeitas a nenhum fator de risco conhecido. Se uma pessoa com câncer de mama tem algum fator de risco, muitas vezes é difícil saber o quanto esse fator pode contribuir para o desenvolvimento da doença.

Fatores de risco que não se pode mudar

Gênero

Ser mulher é o principal fator de risco para o desenvolvimento de câncer de mama.

Idade
O risco aumenta com a idade. Cerca de 12% dos cânceres de mama invasivos são diagnosticados em mulheres com até 45 anos, enquanto cerca de 60% em mulheres acima de 55 anos.

Fatores Genéticos
Cerca de 5 a 10% dos casos de câncer de mama são hereditários. A causa mais comum de câncer de mama hereditário é uma mutação herdada nos genes BRCA1 e BRCA2. Mutações em outros genes, embora raras, podem também levar ao câncer de mama hereditário, como, por exemplo, ATM, TP53, CHEK2 (síndrome de Li-Fraumeni), PTEN (doença de Cowden), CDH1, e STK11 (síndrome de Peutz-Jeghers).

Histórico Familiar
O risco de câncer de mama é maior entre as mulheres com parentes em primeiro grau (mãe, irmã ou filha) que tiveram a doença. Nesses casos o risco da doença praticamente dobra. Ter dois parentes de primeiro grau aumenta o seu risco cerca de 3 vezes.

Histórico Pessoal
Uma mulher com câncer de mama tem um risco de 3 a 4 vezes maior de desenvolver um novo câncer de mama. Isso é diferente de uma recidiva (retorno do tumor).

Raça e Etnia
As mulheres brancas são ligeiramente mais propensas a desenvolver câncer de mama do que as negras. No entanto, em mulheres com menos de 45 anos, o câncer de mama é mais comum em mulheres negras.

Mamas Densas
Mulheres com mamas densas têm um risco aumentado de câncer de mama em relação às mulheres com mamas menos densas. Uma série de fatores pode afetar a densidade da mama, como idade, estado menopausal, uso de medicamentos, gravidez e genética.

Doenças Benignas da Mama
Mulheres diagnosticadas com determinadas condições benignas da mama podem ter um risco aumentado de câncer de mama. As doenças benignas da mama são classificadas de acordo com o risco:

- **Lesões não-proliferativas:** Não estão associadas ao crescimento excessivo do tecido mamário e não parecem afetar o risco de câncer de mama, incluem fibrose e/ou cistos simples, hiperplasia, adenose, ectasia ductal, tumor filoide, papiloma único, necrose, fibrose periductal, metaplasia escamosa e apócrina, calcificações, tumores benignos, como lipoma, hamartoma, hemangioma, neurofibroma e adenomioepitelioma.
- **Lesões proliferativas sem atipia:** Estas condições mostram o crescimento excessivo das células dos ductos ou lobos e incluem hiperplasia ductal, fibroadenoma, adenose esclerosante, papilomatose e cicatriz radial.
- **Lesões proliferativas com atipia:** Nestas condições, existe um crescimento excessivo das células dos ductos ou lobos, com algumas das células normais não aparecendo. Eles têm um forte efeito sobre o risco de câncer de mama, elevando-o de 3 a 5 vezes. Estes tipos de lesões incluem: hiperplasia ductal atípica e hiperplasia lobular atípica.

Menstruação
As mulheres que tiveram menarca precoce (antes dos 12 anos) ou tiveram a menopausa após os 55 anos têm um risco aumentado de câncer de mama. O aumento do risco pode ser devido a uma exposição mais longa a hormônios femininos.

Radioterapia Prévia
As mulheres que fizeram radioterapia na região do tórax têm um risco aumentado de câncer de mama.

Exposição ao Dietilestilbestrol
Mulheres grávidas que receberam dietilestilbestrol (DES) têm um risco ligeiramente maior de desenvolver câncer de mama. Mulheres cujas mães tomaram DES durante a gravidez também podem ter um risco maior de câncer de mama.

Fatores relacionados ao estilo de vida

Ter Filhos
As mulheres que não tiveram filhos ou que tiveram o primeiro filho após os 30 anos têm um risco aumentado de câncer de mama. Ter muitas gestações e engravidar jovem reduz o risco de câncer de mama.

Controle da Natalidade
O uso de pílulas anticoncepcionais aumenta o risco de câncer de mama em relação às mulheres que nunca usaram. Esse risco volta ao normal após a interrupção do uso dos contraceptivos. Mulheres que pararam de usar os anticoncepcionais há mais de 10 anos não parecem ter qualquer aumento no risco.

Reposição Hormonal após a Menopausa
O uso de estrogênio sozinho após a menopausa não parece aumentar o risco de câncer de mama.

Amamentação
Alguns estudos sugerem que a amamentação pode diminuir o risco de câncer de mama.

Alcoolismo
O uso de álcool está claramente associado a um aumento do risco de desenvolver câncer de mama. O risco aumenta com a quantidade de álcool consumida.

Obesidade
Estar acima do peso ou ser obesa após a menopausa aumenta o risco de câncer de mama. Mas a ligação entre o peso e o risco da doença é complexa. Por exemplo, o risco parece ser maior em mulheres que ganharam peso na idade adulta, e não para aquelas que sempre estiveram acima do peso desde a infância.

Atividade Física
Crescem as evidências de que a atividade física na forma de exercício reduz o risco de câncer de mama.

Descrição do Problema

O Câncer de Mama possui 5 estágios, que variam de 0 a 4 e possuem as suas devidas subdivisões. Saiba mais sobre cada um deles no quadro abaixo.

Estágio	Caracterização
Estágio 0	As células cancerosas permanecem no interior do ducto mamário, sem invasão do tecido que há em volta.
Estágio 1A	O tumor possui medida de até 2 cm e ainda não se espalhou para fora da mama. Além disso, não há gânglios linfáticos envolvidos.
Estágio 1B	Esse estágio pode se dar de duas maneiras: 1) Ao invés de ter um tumor na mama, há vários pequenos grupos de células cancerosas (que medem entre 0,2 mm a 2 mm) nos gânglios linfáticos. 2) Há tumor na mama, mas não é maior que 2 cm. Além disso, há diversos grupos de células cancerosas localizados nos nódulos linfáticos.
Estágio 2A	O estágio 2A pode se apresentar de 3 maneiras diferentes: 1) Não há tumor, mas células cancerosas são encontradas nos gânglios linfáticos presentes na axila. 2) O tumor pode medir até 2 cm e se espalha para os linfonodos axilares. 3) O tumor é maior do que 2 cm, mas não mais do que 5 cm, e não se espalhou para os gânglios linfáticos da axila.
Estágio 2B	Nesse estágio, o tumor pode se apresentar de 2 formas: 1) Possui um tamanho maior do que 2 cm, mas não mais do que 5 cm, e se espalhou para os gânglios linfáticos axilares. 2) O tumor é maior do que 5 cm, mas não se espalhou para os gânglios linfáticos axilares.
Estágio 3A	Novamente, esse estágio pode ser de 2 tipos: 1) Não há tumor. O câncer é encontrado em nódulos linfáticos da axila que estão rentes a ela ou a outras estruturas, ou, ainda, que estejam perto do esterno. 2) O tumor pode ter qualquer tamanho e ele já se espalhou nos gânglios linfáticos da axila que estão rentes a ela ou a outras estruturas, ou, ainda, que estejam perto do esterno.
Estágio 3B	No estágio 3B, o câncer pode ser de qualquer tamanho e já se espalhou para as paredes do seio e/ou para a pele da mama. Além disso, ele pode ter se espalhado também para os linfonodos axilares ou ter furado outras estruturas. Ainda, o câncer pode ter se espalhado para os gânglios linfáticos que estão perto do esterno.
Estágio 3C	Nesse estágio, pode ser que não haja nenhum sinal de tumor na mama ou pode ser que o tumor tenha se espalhado para a parede do peito e/ou na pele que reveste a mama. Além disso, o câncer pode ter se espalhado também para os gânglios linfáticos acima ou abaixo da clavícula. Ainda, o câncer pode se espalhar para os linfonodos axilares ou para os gânglios linfáticos próximos ao esterno.
Estágio 4	O Câncer se espalhou para outras partes do corpo (acontecimento chamado de metástase).

Segundo a Federação Brasileira de Instituições Filantrópicas de Apoio à Saúde da Mama (Femama), cerca de 95% dos casos de câncer de mama possui chances de cura. Porém, para que isso possa acontecer, é preciso que a doença seja diagnosticada precocemente, ou seja, quando o tumor ainda medir menos de 1 cm.

A mamografia é o único exame capaz de diagnosticar o câncer de mama em seu estágio inicial, pois os nódulos com menos de 1 cm ainda não podem ser apalpados – o que quebra o tabu de que o autoexame é o primeiro exame que precisa ser feito em caso de suspeita de câncer de mama.

Por conta disso, recomenda-se que mulheres acima de 40 anos se consultem regularmente com o seu mastologista – especialista no funcionamento e afecções da mama – e devem realizar, ao menos, uma mamografia ao ano.

Para tratar um câncer de mama, é preciso avaliar o tipo e o estágio em que se encontra. Feito isso, a definição terapêutica pode ser determinada. O tratamento sempre terá o objetivo de cuidar bem do ser humano e ofertar a ele o melhor método disponível para que ele possa ser efetivamente curado, ou, para aqueles casos em que a cura não seja possível, que ela possa viver com dignidade e com qualidade de vida pelo maior tempo possível.

Sendo assim, seria de extrema aplicabilidade que a partir de dados e resultados dos exames dos pacientes, pudéssemos aplicar inteligência artificial (*machine learning*) no diagnóstico desta doença a fim de anteciparmos seu diagnóstico para que os tratamentos possam ser iniciados com o objetivo de retardar o progresso, efeito e até possibilitar o tratamento definitivo.

Análise dos dados

As séries de dados foram obtidas da plataforma Kaggle[2] e contém um conjunto de dados para ser utilizado como treinamento, validação e testes. Para isso os dados foram separados utilizando um método aleatório para criar cada um dos subconjuntos de dados 80%, 10% e 10% respectivamente.

Com isso, a composição dos dados ficou da seguinte estrutura e tipos de informação:

Tipo	%	Quantidade
Dados de Treinamento	80	455
Dados de teste	10	57
Dados de Validação Cruzada	10	57
Total de registros	100	569

Tabela 01: Composição dos dados por tipo

1) Número de Identificação (ID number)

2) Diagnóstico (Diagnosis (M = malignant, B = benign))

3-32) Dez valores reais computados para cada núcleo celular:

a) Raio (media da distância do centro aos pontos do perímetro)
b) Textura (desvio padrão dos valores na escala de cinza)
c) Perímetro
d) Área
e) Suavidade (variação local no comprimento do raio)
f) Compactação (perímetro ^2 / área – 1.0)
g) Concavidade (gravidade das porções côncavas do contorno)
h) Pontos côncavos (número de porções côncavas do contorno)
i) Simetria
j) Dimensão fractal (“aproximação litorânea” -1

Sendo que a classificação está distribuída conforme figura 01.

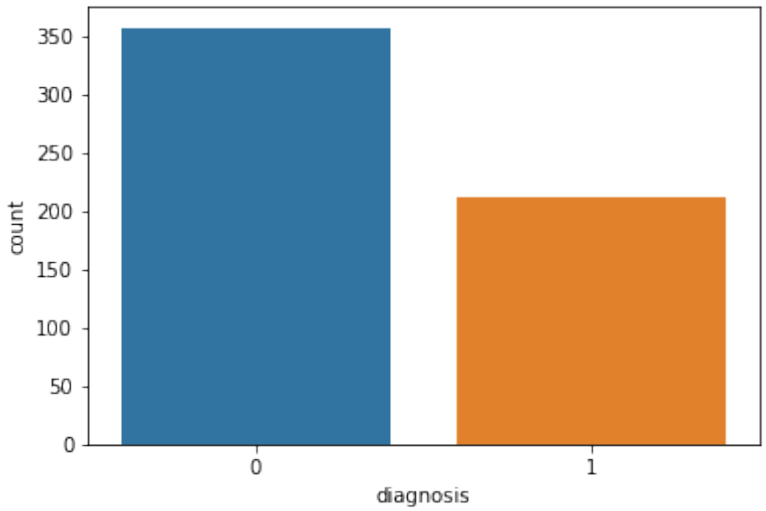

Figura 01: Gráfico em barras da quantidade de diagnósticos Malígnos e Benígnos

Tratamento e Modelagem dos dados

O valor estandardizado de uma distribuição de dados é também chamado de z-score ou valor transformado. A estandardização é diferente da normalização dos dados, onde se objetiva que cada variável se encontre em uma faixa de valores no intervalo [0,1]. Na estandardização a faixa de valores pode variar e depende do desvio padrão e se utiliza a fórmula abaixo:

$$x_{estand} = \frac{x - \bar{x}}{\sigma_x}$$

Calculada, como na normalização, de forma independente para cada variável. A estandardização toma informações sobre a média e o desvio padrão de uma variável e produz um valor correspondente a cada valor original que especifica a posição deste valor original dentro da distribuição original de dados.

Uma vez executada a estandardização, foi utilizado a biblioteca de visualização de dados seaborn, sendo que em nossa análise utilizamos a função sns.swarmplot[5], conforme ilustrado nas figuras abaixo, nas quais é possível observar padrões de classificação entre cada uma das features e seus respectivos diagnósticos (figuras 03 e 04)

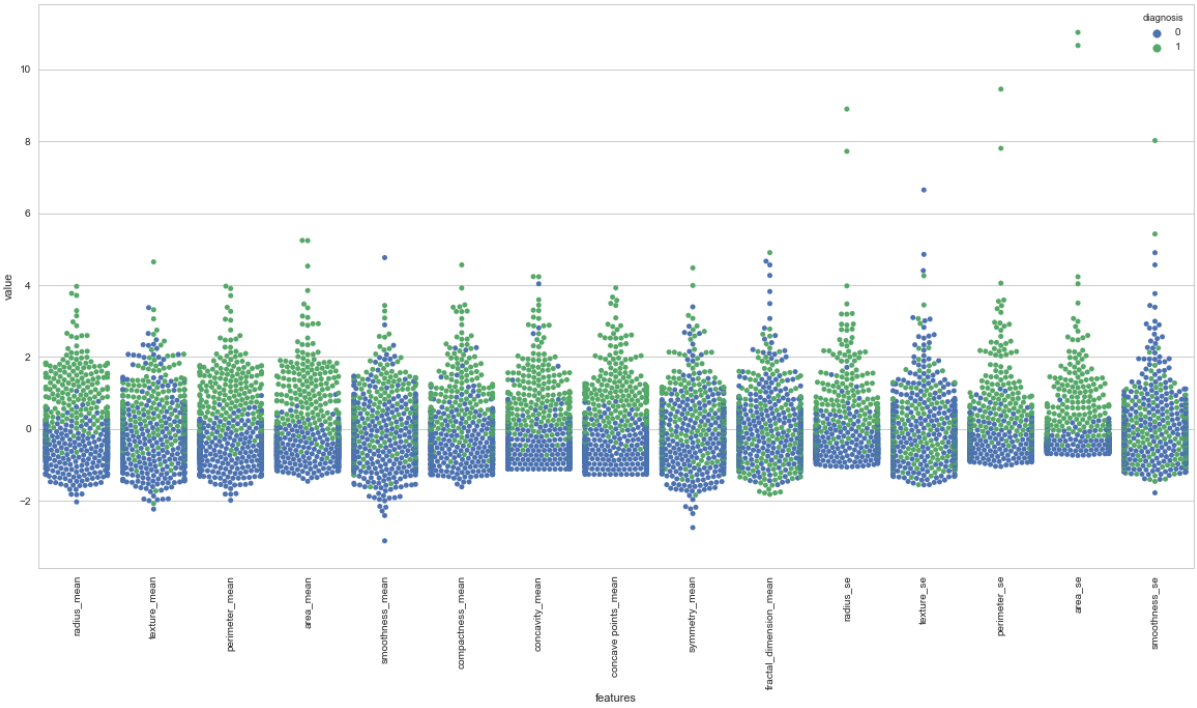

Figura 02: Padrão de classificação por feature (parte 1)

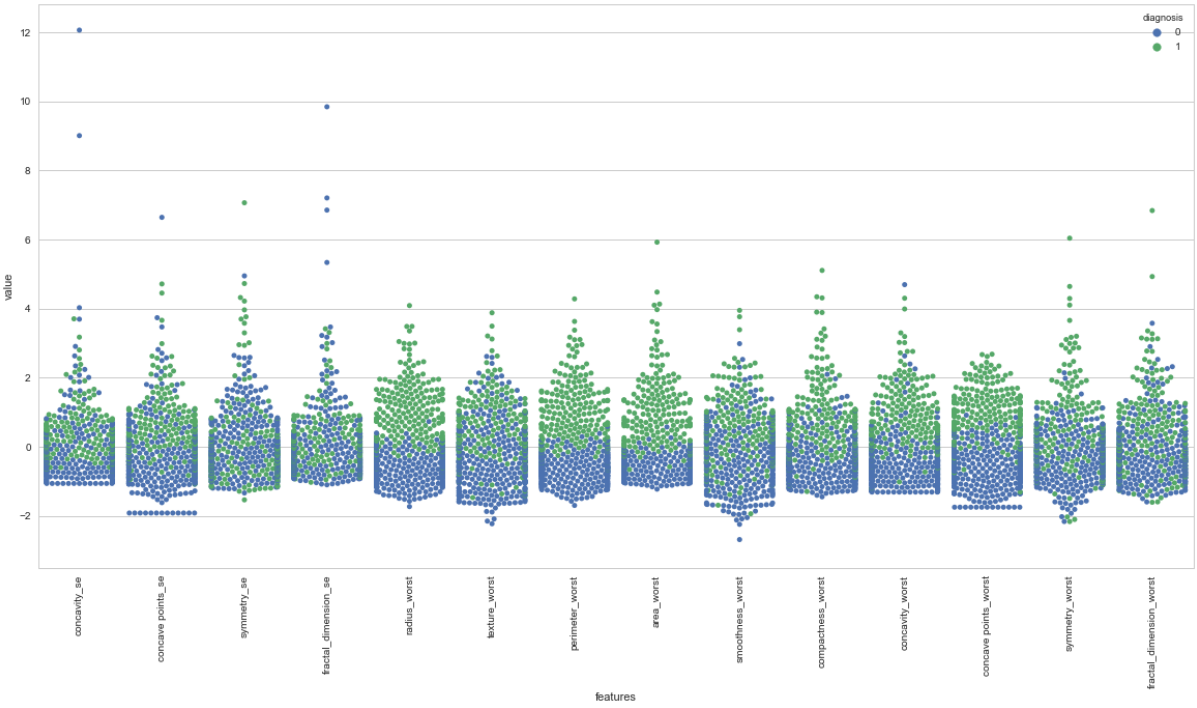

Figura 03: Padrão de classificação por feature (parte 2)

A partir destas visualizações iniciais, foram aplicados sobre os dados alguns algoritmos para determinar, para um modelo genérico (RandomForestClassifier), quais as features mais relevantes para os dados propostos. Para isso, utilizou-se as seguintes funções e obtivemos os resultados apresentados na tabela 02.

- SelectKBest: Seleciona as features de acordo com o K maiores placares (scores)
- RFECV: Ranqueamento das features com eliminação de features recursivas e seleção de validação cruzada das melhores features.

Seleção de Feature	Quantidade de Features	Precisão
N/A	30	0.92982
SelectKBest(chi2, k=5)	5	0.9649
RFECV(estimator=clf3, step=1, cv=5, scoring='accuracy')	20	0.94736

Tabela 02: Correlacionamento dos dados e impacto na precisão

A partir destes resultados, vemos que conseguimos obter uma melhor precisão quando aplicamos a função SelectKBest, entretanto, também algumas variações foram percebidas para diferentes algoritmos, como é possível visualizar nas figuras abaixo:

LogisticRegression	AdaBoostClassifier	DecisionTreeClassifier

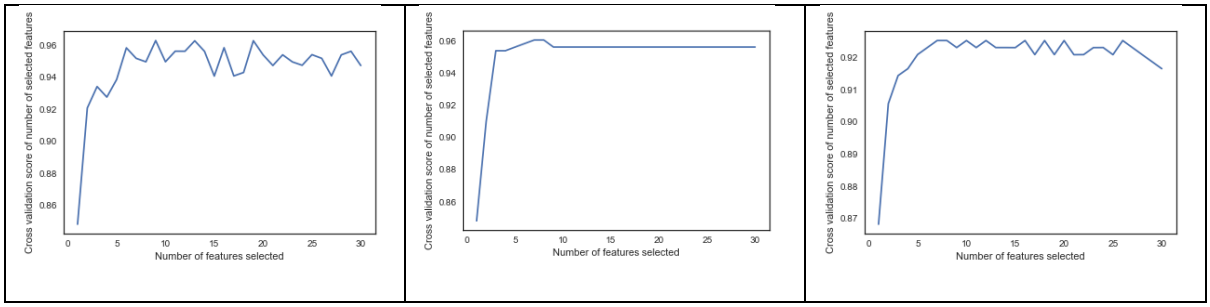

Figura 04: Comparativos de precisão x # de features para diferentes algoritmos de classificação

Neste caso, não aplicarei nenhuma redução no número de features disponíveis e deixarei a cargo do algoritmo encontrar a melhor modelagem, pois não identifiquei ganhos significativos no resultado da Precisão ou velocidade de processamento para o tamanho da amostra de disponível.

Métricas de avaliação

Antes de definirmos quais métricas serão avaliadas, é importante entende-las para que possamos assertivamente observar seus resultados e decidir suas relevâncias para os modelos e problema.

Accuracy: é a medida de desempenho mais intuitiva e é simplesmente uma proporção da observação corretamente prevista para o total de observações. Pode-se pensar que, se tivermos alta precisão, nosso modelo é o melhor. Sim, a precisão é uma ótima medida, mas apenas quando você tem conjuntos de dados simétricos em que os valores de falso positivo e falso negativo são quase iguais. Portanto, você precisa examinar outros parâmetros para avaliar o desempenho do seu modelo.

$$Accuracy = \frac{t_p + t_n}{t_p + f_p + t_n + f_n}$$

Precision: é a razão de observações positivas preditas corretamente para o total de observações positivas previstas. A pergunta que esta resposta métrica é de todos os passageiros que rotularam como sobrevivido, quantos realmente sobreviveram? A alta precisão está relacionada à baixa taxa de falsos positivos.

$$Precision = \frac{t_p}{t_p + f_p}$$

Recall: é a proporção de observações positivas preditas corretamente para todas as observações na classe real.

$$Recall = \frac{t_p}{t_p + f_n}$$

F1-Score: é a média ponderada de Precision e Recall. Portanto, essa pontuação leva em conta tanto os falsos positivos quanto os falsos negativos. Intuitivamente, não é tão fácil entender como precisão, mas F1 é geralmente mais útil que precisão, especialmente se você tiver uma distribuição de classe irregular. A precisão funciona melhor se falsos positivos e falsos negativos tiverem um custo similar. Se o custo de falsos positivos e falsos negativos for muito diferente, é melhor analisar tanto o Precision quanto o Recall

$$F1\ Score = 2\frac{Recall * Precision}{Recall + Precision}$$

R2-Score: é uma medida estatística de quão próximos os dados estão da linha de regressão ajustada. É também conhecido como o coeficiente de determinação, ou o coeficiente de determinação múltipla para regressão múltipla. A definição de R-quadrado é bastante direta; é a porcentagem da variação da variável resposta que é explicada por um modelo linear.

Modelo de referência (benchmark)

O modelo de referência deste projeto é o utilizado por Silvia, Nadia et all no estudo realizado com o objetivo de desenvolvimento de um sistema inteligente composto de Redes Neurais Artificiais (RNAs) do tipo MultiLayer Perceptron (MLP)aplicado ao problema de diagnóstico de câncer de mama.

O resultado obtido neste estudo foi bastante expressivo onde verificou-se que o modelo proposto obteve uma taxa de acerto de 95.32%, sendo uma opção válida para um diagnóstico preciso e eficaz de câncer de mama.

Neste caso, espera-se que os estudos e resultados deste projeto alcance valores da taxa de acerto (precisão) próximos ao resultado obtivo por Silvia, Nadia, conforme figura abaixo [4].

Tabela 1: Resultados obtidos - conjunto de teste.

	B	M	Acerto (%) por Classe
B	103	8	92.79
M	0	60	100
Total	103	68	95.32

Figura 05: Tabela com resultado obtido no estudo [4].

Descrição da Solução e Justificativa

Baseado no modelo de referência (benchmark) proposto por Silvia, Nadia et all, eu simulei alguns algoritmos disponíveis na biblioteca do scikit-learn para determinar um conjunto possível de soluções os quais pudessem resultar em soluções próximas ou melhores que a encontrada neste estudo de referência.

Para isso, apliquei uma rotina em python para que simulasse para uma série de algoritmos as mesmas condições de dados e parâmetros cujos resultados são apresentados na tabela Eu estou buscando por algoritmos que permitirão analisar amostras dentre uma categoria para que realize uma classificação binária. Dentre as possíveis opções, irei avaliar em princípio dois classificadores:

Algoritmo	Training size	F1-Score Training	F1-Score Testing
LogisticRegression	200	0,9466	0,9091
	450	0,9426	0,9286
KNeighborsClassifier	200	0,9355	0,8679
	450	0,9161	0,8889
AdaBoostClassifier	200	1,0000	0,9310
	450	0,9762	0,9655
NearestCentroid	200	0,8276	0,8889

	450	0,8138	0,8889
GaussianProcessClassifier	200	1,0000	0,4500
	450	1,0000	0,5714
DecisionTreeClassifier	200	0,9489	0,8966
	450	0,9649	0,8966
RandomForestClassifier	200	0,9767	0,9286
	450	0,9820	0,9123
GradientBoostingClassifier	200	1,0000	0,9153
	450	1,0000	0,9310

Tabela 03: Métrica F1 Score aplicados à diversos algoritmos de classificação

Como é possível observar pelos resultados, vemos 4 algoritmos com resultado do F1-Score acima de 90%:

- LogisticRegression
- AdaBoostClassifier
- RandomForestClassifier
- GradientBoostingClassifier

Neste caso, irei considerar o algoritmo AdaBoostClassifier por ter apresentado melhor métrica entre todos os algoritmos testados.

Refinamento

AdaBoostClassifier é um meta-estimador que começa ajustando um classificador no conjunto de dados original e então ajusta cópias adicionais do classificador no mesmo conjunto de dados, mas onde os pesos das instâncias classificadas incorretamente são ajustados de forma que os classificadores subsequentes se concentrem mais casos difíceis.

Este modelo implementa o algoritmo conhecido como AdaBoost-SAMME[6].

Inicialmente, apliquei o algoritmo com os parâmetros padrões da biblioteca e em seguida alterando alguns parâmetros de modo a buscar um melhor resultado.

Algoritmo	Dados	Accuracy	R2 Score	Cross-Predicted	Precision	Recall	F1-Score	Time (seg)
Default	Treinamento	0,9691	1,0000	0,9691	0,9826	0,9342	1,0000	11,66
	Teste	0,9474	0,7889	0,8800	-	-	0,9434	
	Cross-Validação	1,0000	1,0000	0,9690	-	-	1,0000	
n_estimators=20	Treinamento	0,9713	0,9906	0,9713	0,9832	0,9404	0,9970	5,00
	Teste	0,9474	0,7889	0,8967	-	-	0,9412	
	Cross-Validação	0,9825	0,9131	0,9690	-	-	0,9697	
n_estimators=100	Treinamento	0,9712	1,0000	0,9712	0,9826	0,9401	1,0000	23,74
	Teste	0,9474	0,7889	0,8967	-	-	0,9434	
	Cross-Validação	1,0000	1,0000	0,9857	-	-	1,0000	
DecisionTreeClassifier(max_depth=1), algorithm="SAMME", n_estimators=100	Treinamento	0,9757	1,0000	0,9757	0,9895	0,9463	1,0000	23,41
	Teste	0,9474	0,7889	0,8733	-	-	0,9434	

	Cross-Validação	1,0000	1,0000	0,9690	-	-	1,0000	
DecisionTreeClassifier(max_depth=5), algorithm="SAMME", n_estimators=100	Treinamento	0,9648	1,0000	0,9604	0,9583	0,9408	1,0000	13,49
	Teste	0,9123	0,6481	0,8800	-	-	0,9057	
	Cross-Validação	0,9825	0,9131	0,9657	-	-	0,9697	

Tabela 04: Resultados obtidos durante o refinamento do modelo.

Conclusão e Pontos de Atenção

A conclusão que chego deste projeto é que conseguimos encontrar um modelo que conseguiu um resultado bastante elevado para o problema proposto > 94% de *Accuracy* para diversos valores de seus parâmetros, o qual podemos considerar um modelo bastante confiável para acelerar diagnósticos de câncer de mama e com isso melhorarmos a assertividade e rapidez nestes procedimentos.

Os pontos de atenção para este projeto é o de conseguirmos uma massa maior de dados para que haja um aumento na confiabilidade dos modelos e resultados. Isso porque em alguns cenários testados vimos algumas métricas atingindo 100% o qual pode induzir a falsos positivos ou negativos.

Vale também avaliar outros modelos para treinamento e testes, inclusive modelos mais leves que possam ser aplicados em equipamentos com menor poder computacional. Outro item que merece bastante atenção é com relação as análises de correlacionamento dos dados, os quais não encontrei grandes motivadores de buscar alternativas para reduzir ou eliminar features que não tivessem representatividade com o diagnóstico.

Sendo assim, quero agradecer a oportunidade deste projeto em poder aplicar os conhecimentos adquiridos durante o curso e ainda aplicá-los a um caso que pode trazer grandes benefícios às pessoas de maneira geral.

Referências

[1] https://minutosaudavel.com.br/cancer-de-mama-sintomas-tipos-cura-o-que-e-prevencao-e-mais/
[2] https://www.kaggle.com/uciml/breast-cancer-wisconsin-data
[3] https://pt.wikipedia.org/wiki/Regressão_log%C3%ADstica
[4] Silvia, Nadia; Salvador, Victor; Cordeiro, Jose; Araujo, Ricardo - Desenvolvimento de um Sistema Inteligente Aplicado ao Diagnóstico de Câncer de Mama via Redes Neurais Artificiais
[5] https://seaborn.pydata.org/examples/scatterplot_categorical.html
[6] J Zhu, H. Zou, S. Rosset, T. Hastie, "Multi-class AdaBoost", 2009.