Aplicação de Machine Learning na Previsão do Cancro da Mama

Catarina Pimenta[1]
Universidade de Coimbra
[1]Estudante,Departamento de Engenharia Informática,
Faculdade de Ciências e Tecnologia

Resumo

O cancro da mama continua a ser uma das principais causas de morte em mulheres e está a aumentar nos países em desenvolvimento, onde a maioria dos casos é diagnosticada em estágios avançados. A deteção e classificação do cancro da mama nas fases iniciais do seu desenvolvimento podem permitir que os pacientes recebam o tratamento adequado. Este artigo analisa o desempenho de sete algoritmos de Machine Learning (ML): : Logistic Regression (LR), Naïve Bayes (NB), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree (DT), Random Forest (RF), Artificial Neural Network (ANN) no Conjunto de Dados de Cancro da Mama de Wisconsin (WBCD) do repositório de ML da Universidade da Califórnia em Irvine (UCI). Para a implementação dos algoritmos de ML, o conjunto de dados foi dividido em fases de treino e teste. Os resultados experimentais mostram que os algoritmos e técnicas que utilizei alcançaram elevado desempenho no diagnóstico do cancro da mama.

Palavras-chave: Classificação do cancro da mama, , Machine Learning, Logistic Regression, Naïve Bayes, Support Vector Machine, K-Nearest Neighbors, Decision Tree, Random Tree, Random Forest, Artificial Neural Network

1 Introdução

O cancro da mama é o cancro mais diagnosticado em mulheres e representa 1 em cada 4 casos anuais de cancro em todo o mundo [1]. Em 2020, a incidência na região da OMS na Europa foi estimada em 576.300 [2] e, na UE-27, foi de 355.500 [3]. Estima-se que 21% dos casos de cancro da mama na Europa ocorram em mulheres com menos de 50 anos, enquanto 35% ocorrem entre os 50 e os 64 anos, e os restantes casos afetam mulheres acima dessa faixa etária [2]. 1 em cada 11 mulheres na UE-27 desenvolverá cancro da mama antes dos 74 anos [3].

O cancro da mama é a principal causa de morte por cancro em mulheres em todo o mundo, representando 1 em 6 mortes por cancro [1], e tira mais vidas de mulheres europeias do que qualquer outro tipo de cancro [2]. Em 2020, o número estimado de mulheres que morreram de cancro da mama na região da OMS na Europa foi de 157.100 [2] e, na UE-27, foi de 91.826 [3].

O aumento do número de casos de cancro da mama pode dever-se a mudanças nos hábitos de vida, como o sedentarismo, ganho de peso, obesidade e consumo de álcool. Fatores reprodutivos, como ter filhos em idade mais jovem (abaixo dos 30 anos), ter vários filhos e amamentar por longos períodos, reduzem o risco de cancro da mama. O uso de contraceptivos orais combinados de estrogénio-progestogénio está associado a um aumento do risco de cancro da mama, especialmente entre mulheres jovens. Mulheres que menstruaram antes dos 11 ou 12 anos ou que passaram pela menopausa após os 55 têm um risco ligeiramente maior de cancro da mama [4].

Este artigo propõe um método de classificação para a classificação de tumores de mama usando métodos de classificação de ML. Avaliámos o desempenho dos seguintes algoritmos de classificação: : Logistic Regression (LR), Naïve Bayes (NB), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree (DT), Random Forest (RF), Artificial Neural Network (ANN). Esses algoritmos proporcionam diferentes valores de precisão. A fonte dos dados é o Conjunto de Dados de Cancro da Mama de Wisconsin (WBCD) retirado do repositório de ML da Universidade da Califórnia em Irvine (UCI).

O objetivo deste estudo é detetar o cancro da mama a partir de um conjunto de dados de um programa de ML em Python.

O restante do artigo está organizado da seguinte forma: a Secção 2 revê trabalhos existentes sobre a deteção de tumores de cancro da mama. A Secção 3 apresenta o material utilizado. A Secção 4 trata da correlação dos dados. A Secção 5 explica a metodologia proposta e os experimentos. A Secção 6 apresenta e discute os resultados obtidos nos experimentos. A Secção 7 conclui o artigo e sugere algumas direções futuras de pesquisa para a classificação de tumores de cancro da mama.

2 Trabalhos Relacionados

A pesquisa usando classificadores de Machine Learning no domínio médico tem sido prevalente há muito tempo, especialmente no diagnóstico de cancro da mama. A tarefa de classificação é um dos tipos populares de tarefas de Machine Learning. Vários estudos de pesquisa foram conduzidos aplicando classificadores de ML a diferentes dados médicos, incluindo o WBCD.

A referência [5] examinou a robustez do Support Vector Machine (SVM) de mínimos quadrados usando precisão de

classificação, análise de sensibilidade e especificidade, método de validação cruzada k-fold e matriz de confusão. Eles obtiveram uma precisão de classificação de 98,53%.

A referência [6] propôs o uso de Redes Neurais Artificiais de Feed Forward e algoritmo de retropropagação para treinar a rede. O desempenho da rede é avaliado usando WBDS para vários algoritmos de treino. A maior precisão de 99,28% é alcançada ao usar o algoritmo de Levenberg Marquardt.

A referência [7] investigou uma nova abordagem de classificação para deteção de anormalidades mamárias em mamografias digitais usando a Rede Neural de Otimização por Enxame de Partículas (PSOWNN). O algoritmo de deteção de anormalidades proposto baseia-se na extração de medidas de energia de textura de Leis de mamogramas e na classificação de regiões suspeitas aplicando um classificador de padrões. Eles alcançaram 93,671%, 92,105% e 94,167% para precisão, especificidade e sensibilidade, respetivamente.

A referência [8] usou um método de relação de indiscernibilidade do conjunto aproximado com uma rede neural de retropropagação (RS-BPNN). Este trabalho tem duas etapas. A primeira etapa trata dos valores em falta para obter um conjunto de dados suave e selecionar atributos apropriados do conjunto de dados clínico pelo método de relação de indiscernibilidade. A segunda etapa é a classificação usando a rede neural de retropropagação. A precisão obtida a partir do método proposto foi de 98,6% no conjunto de dados de cancro da mama.

Neste estudo, comparei sete algoritmos de ML: LR, NB, SVM, KNN, DT, RF, ANN, e as suas correlações, também comparando com o atributo previsto (saída). O objetivo é fornecer o melhor conjunto de características para aumentar a precisão e o desempenho do diagnóstico do cancro da mama.

3 Materiais

3.1 Conjunto de dados

Neste estudo, utilizei o Conjunto de Dados Público de Cancro da Mama de Wisconsin (WBCD) do repositório de Machine Learning da UCI, que fornece informações muito sólidas e úteis com o único propósito de construir modelos para prever se o cancro é benigno ou maligno. O conjunto de dados com o qual vou trabalhar possui 32 atributos (o atributo 'ID' é omitido) e 568 instâncias. De todos os atributos, apenas 1 é categórico, como mostrado na seguinte figura.

ATRIBUTE NAME	DESCRIPTION	TYPE	DOMAIN
DIAGNOSIS	The diagnosis of breast tissues	Categorical	M or B (Malignant or Benign)
RADIUS_MEAN	Mean of distances from center to points on the perimeter	Numeric	From 6.98 to 28.1
TEXTURE_MEAN	Standard deviation of gray-scale values	Numeric	From 9.71 to 39.3
PERIMETER_MEAN	Mean size of the core tumor	Numeric	From 43.8 to 189
AREA_MEAN	Mean area size for the core tumor	Numeric	From 144 to 2500
SMOOTHNESS_MEAN	Mean of local variation in radius lengths	Numeric	From 0.05 to 0.16
COMPACTNESS_MEAN	Mean of perimeter^2 / area - 1.0	Numeric	From 0.02 to 0.35
CONCATIVITY_MEAN	Mean of severity of concave portions of the contour	Numeric	From 0 to 0.43
CONCAVE POINTS_MEAN	Mean for the number of concave portions of the contour	Numeric	From 0 to 0.2
SIMETRY_MEAN	Mean symmetry for the core tumor	Numeric	From 0.11 to 0.3
FRACTAL_DIMENSION_MEAN	Mean for "coastline approximation" – 1.0	Numeric	From 0.05 to 0.1
RADIUS_SE	Standard error for the mean of distances from center to points on the perimeter	Numeric	From 0.11 to 2.87
TEXTURE_SE	Standard error for the standard deviation of gray-scale values	Numeric	From 0.36 to 4.88
PERIMETER_SE	Standard error for the size of the core tumor	Numeric	From 0.76 to 22
AREA_SE	Standard error for the area size for the core tumor	Numeric	From 6.8 to 542
SMOOTHNESS_SE	Standard error for local variation in radius lengths	Numeric	From 0 to 0.03
COMPACTNESS_SE	Standard error for perimeter^2 / area – 1.0	Numeric	From 0 to 0.14
CONCAVITY_SE	Standard error for severity of concave portions of the contour	Numeric	From 0 to 0.4
CONCAVE POINTS_SE	Standard error for the number of concave portions of the contour	Numeric	From 0 to 0.05
SYMMETRY_SE	Standard error for the symmetry size for the core tumor	Numeric	From 0.01 to 0.08
FRACTAL_DIMENSION_SE	Standard error for "coastline approximation" – 1	Numeric	From 0 to 0.03
RADIUS_WORST	Worst (or largest) value for mean of distances from center to points on the perimeter	Numeric	From.93 to 36
TEXTURE_WORST	Worst (or largest) value for standard deviation of gray-scale values	Numeric	From 12 to 49.5
PERIMETER_WORST	Worst (or largest) value for the size of the core tumor	Numeric	From 50.4 to 251
AREA_WORST	Worst (or largest) value for the area size for the core tumor	Numeric	From 185 to 4250
SMOOTHNESS_WORST	Worst (or largest) value for local variation in radius lengths	Numeric	From 0.07 to 0.22
COMPACTNESS_WORST	Worst (or largest) value for perimeter^2 / area – 1.0	Numeric	From 0.03 to 1.06
CONCATIVITY_WORST	Worst (or largest) value for severity of concave portions of the contour	Numeric	From 0 to 1.25
CONCAVE POINTS_WORST	Worst (or largest) value for number of concave portions of the contour	Numeric	From 0 to 0.29
SYMETRY_WORST	Worst (or largest) value for the symmetry size for the core tumor	Numeric	From 0.16 to 0.66
FRACTAL_DIMENSION_WORST	Worst (or largest) value for "coastline approximation" –1	Numeric	From 0.06 to 0.21

Figura 1. Conjunto de Dados
Nota: É necessário ampliar a imagem para visualizar
Também representei graficamente os dados de diagnóstico para ter uma boa perceção de quantos desses resultados eram Malignos ou Benignos.

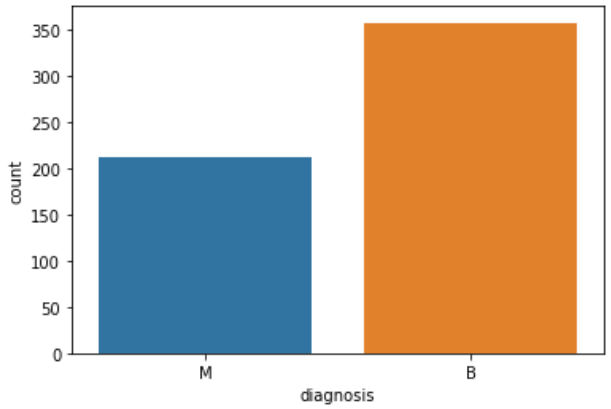

Figura 2. Número de resultados Malignos ou Benignos.

3.2 Frameworks

Para o código implementado neste estudo, foram aplicadas múltiplas bibliotecas e métodos a fim de testar várias alternativas para resolver o problema. Para a maioria destas, utilizei a biblioteca conhecida como Sklearn, que apresenta vários algoritmos de classificação, regressão e clustering, entre os quais estão o K-Nearest Neighbor, Gaussian's Naive Bayes, Support Vector Classifier, Logistic Regression, Decision Tree, Random Forest. Foi também utilizado uma Rede Neural, o que exigiu a utilização da biblioteca Keras. As outras bibliotecas utilizadas, como seaborn, matplotlib, são usadas para analisar dados em gráficos e plots, e a biblioteca Pandas foi utilizada para extrair o conjunto de dados.

4 Correlação dos Dados

4.1 Visualizar Dispersão e Relação de Variáveis

É importante conhecer a dispersão dos valores de diagnóstico para cada variável. Para isso, utilizei um pairplot do seaborn com as colunas do meio e a coluna de diagnóstico para entender e visualizar melhor a relação entre as variáveis e a dispersão do diagnóstico em cada relação. Não utilizei todas as variáveis porque seria difícil visualizar, portanto, optei por representar os valores médios.

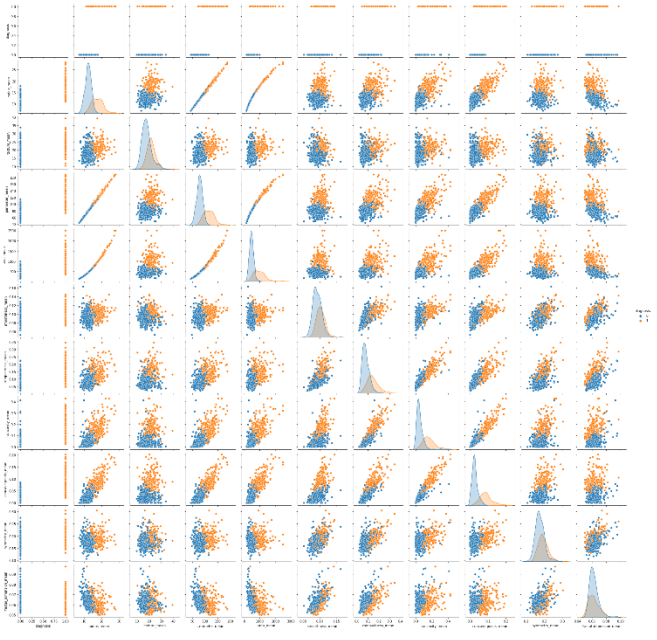

Figura 3. Gráfico de Relação e Dispersão
Nota: É necessário ampliar a imagem para visualizar

No gráfico acima, podemos observar que existem relações de variáveis cujo diagnóstico não é disperso, como na relação entre texture_mean e fractal_dimension_mean, e outras em que o oposto é verdadeiro, como na relação entre texture_mean e radius_mean.

4.2 Análise de Características (Heat Map)

O heat map representa as correlações entre cada variável através de um esquema de cores, sendo quanto mais escuro o tom, mais forte é a correlação entre cada variável e quanto mais claro, mais fraca é.

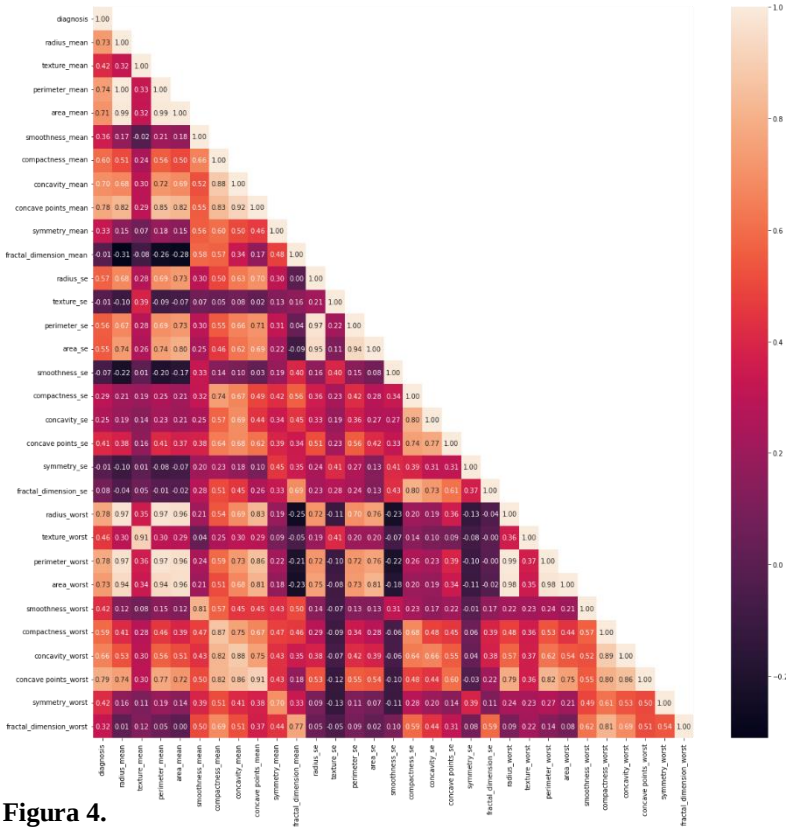

Figura 4.
Heat Map de todas as correlações entre colunas.
Nota: É necessário ampliar a imagem para visualizar os valores

Para a previsão do cancro da mama, apenas a correlação da coluna de decisão com as outras colunas é relevante para o resultado, assim a correlação entre cada variável é ignorada.

Se analisarmos profundamente, é possível concluir que a correlação mais alta é 0,79 (entre concave point worst e diagnóstico), sendo uma correlação forte. Com isso, podemos concluir que as restantes variáveis terão correlações moderadas, fracas ou nulas.

Considerando estes resultados e o objetivo deste projeto, nenhuma das correlações no mapa de calor pode ser excluída da pesquisa, porque, apesar de pequenas, as correlações acabam por afetar, mesmo que minimamente, os resultados.

4.3 Conclusão da Correlação

Após o estudo e análise do heat map, no projeto irei utilizar as seguintes variáveis com as correspondentes correlações (entre a variável e diagnóstico):

Radius mean: 0,73
Texture_mean: 0,42
Perimeter_mean: 0,74
Area_mean: 0,71
Smoothness_mean: 0,36
Compactness_mean: 0,60
Concavity_mean: 0,70
Concave_points_mean: 0,78
Symmetry_mean: 0,33
Fractal_dimension_mean: -0,01
Radius_se: 0,57
Texture_se: -0,01
Perimeter_se: 0,56
Area_se: 0,55
Smoothness_se: -0,07
Compactness_se: 0,29
Concavity_se: 0,25
Concave_points_se: 0,41
Symmetry_se: -0,01
Fractal_dimension_se: 0,08
Radius_worst: 0,78
Texture_worst: 0,46
Perimeter_worst: 0,78
Area_worst: 0,73
Smoothness_worst: 0,42
Compactness_worst: 0,59
Concavity worst: 0,66
Concave_points_worst: 0,79
Symmetry_worst: 0,42
Fractal_dimension_worst: 0,32

5 Métodos

5.1 Metodologia

O objetivo deste estudo é identificar, dentro do conjunto de dados, quais são os fatores mais importantes e as correlações entre eles para obter um modelo que preveja com precisão se alguém tem a doença ou não. Iniciei importando/extraindo o conjunto de dados com 33 atributos/colunas e 569 instâncias/linhas de informação, incluindo a coluna sobre o diagnóstico. Em seguida, procedi à limpeza dos dados, verificando valores nulos e removendo-os, além de verificar o tipo das variáveis para garantir que estamos a trabalhar com o esperado. Através do Pandas, consegui ter uma melhor visualização e compreensão de cada atributo e os seus valores.

Também utilizei um pairplot do seaborn com as colunas de média e a coluna de diagnóstico para entender e visualizar melhor a relação de cada variável com o diagnóstico. Não utilizei todas as variáveis porque tornaria difícil visualizar, por isso, optei por representar apenas os valores médios de cada variável.

Para analisar quais atributos vou utilizar nos métodos, desenhei um heat map da correlação entre todas as colunas, considerando apenas a correlação entre colunas e diagnóstico como importante. Pois é essa relação que ajudará a obter melhores resultados. O heat map levou-me a concluir que todas as variáveis são importantes para obter um resultado melhor, independentemente da força da correlação.

O conjunto de dados foi então dividido em 55% e 45% para treino e teste do modelo, respetivamente. A precisão/eficiência do modelo é obtida com o conjunto de dados de teste. Para construir este modelo, utilizei vários algoritmos do sklearn, tais como:

Os algoritmos:

1. *Logistic Regression*

O modelo de Regressão Logística mede, através de probabilidades, a relação entre uma variável dependente categórica e todas as variáveis independentes, sendo a variável dependente o diagnóstico, e todas as outras são as variáveis independentes [9].

2. *Naïve Bayes*

Com base no teorema de Bayes, o Naïve Bayes é um algoritmo probabilístico de aprendizagem que calcula a probabilidade de um evento ocorrer dado que outro evento ocorreu, por pressuposição [10].

3. *Support Vector Machine (SVM)*

O propósito de um algoritmo SVM é encontrar um hiperplano num espaço N-dimensional que classifica distintamente os pontos de dados [11].

4. *K-Nearest Neighbors*

O propósito do algoritmo KNN é avaliar a classe correta para os pontos de dados de teste, calculando a distância entre estes e os pontos de dados de treino [12].

5. *Decision Tree*

Uma Decison Tree pode ser usada para representar visualmente e explicitamente decisões e tomada de decisões. Como o nome sugere, utiliza um modelo em forma de árvore para decisões. Apesar de ser uma ferramenta comumente utilizada em mineração de dados para derivar uma estratégia para alcançar um objetivo específico, é também amplamente usada em machine learning [13].

6. *Random Forest*

Uma Forest Tree consiste em muitas Decision Trees que operam como um conjunto, cada Decision Tree individual dá uma saída e a saída com mais votos torna-se a previsão deste modelo [14].

7. *Artificial Neural Network*

A sua estrutura é formada por camadas de nós, contendo uma camada de entrada, uma ou mais camadas ocultas e uma camada de saída. Cada nó está conectado a outro nó e possui um peso e um limiar associados. Se a saída de qualquer nó individual estiver acima do valor de limiar especificado, o nó envia dados para a próxima camada da rede, caso contrário, nenhum dado é transmitido [15].

Medi a precisão de cada modelo de classificação para avaliar o desempenho de cada um. A precisão do modelo é a fração de todo o conjunto de dados de teste que o modelo previu corretamente.

5.2 Experimentos

Os experimentos neste estudo baseiam-se na seleção de características através do desempenho dos modelos de classificação implementados, e optei por apresentar como resultado aquele com o melhor resultado. Essa avaliação, como mencionado anteriormente, baseia-se na precisão(accuracy) de cada modelo medida.

6 Resultados e Discussão

Nas etapas anteriores, foram feitas várias seleções e escolhas sobre como usar o conjunto de dados para obter um resultado melhor, com maior precisão.

A partir dos métodos mencionados acima e das variáveis selecionadas a partir das correlações, foi possível treinar os métodos para uma certa aleatoriedade de 55% dos dados e alcançar a sua *Training Accuracy*:

Accuracy dos modelos:

[0] Neural Network Training Accuracy:
0.9968
[1]K Nearest Neighbor Training Accuracy:
0.9807692307692307
[2] Gaussian Naive Bayes Training Accuracy:
0.9519230769230769
[3]SVC Linear Training Accuracy:
0.9903846153846154
[4]SVC RBF Training Accuracy:
0.9807692307692307
[5] Logistic Regression Training Accuracy:
0.9903846153846154
[6] Decision Tree Classifier Training Accuracy:
1.0
[7] Random Forest Classifier Training Accuracy:
0.9967948717948718

Após ter sido treinado, consegui finalmente testar os restantes 45% dos dados nos métodos e obter a precisão(accuracy) de cada um, conforme mostrado abaixo:

Accuracy dos modelos:

[0] Neural Network Testing Accuracy:
0.9689
[1]K Nearest Neighbor Testing Accuracy:
0.9494163424124513
[2] Gaussian Naive Bayes Testing Accuracy:
0.9377431906614786
[3]SVC Linear Testing Accuracy:
0.9571984435797666

[4]SVC RBF Testing Accuracy:
0.980544747081712
[5] Logistic Regression Testing Accuracy:
0.9649805447470817
[6] Decision Tree Classifier Testing Accuracy:
0.914396887159533
[7] Random Forest Classifier Testing Accuracy:
0.9416342412451362

Através da precisão(accuracy) calculada, conclui-se que o melhor resultado é obtido utilizando o método SVC RBF:

Model SVC rbf

	precision	recall	f1-score	support
0	0.98	0.99	0.98	160
1	0.99	0.96	0.97	97
accuracy			0.98	257
macro avg	0.98	0.98	0.98	257
weighted avg	0.98	0.98	0.98	257

Figura 5. Modelo SVC RBF

Este método fornece os melhores resultados deste trabalho, mas ainda tem falhas. Se compararmos os resultados previstos pelo método com os resultados reais, podemos observar o seguinte gráfico:

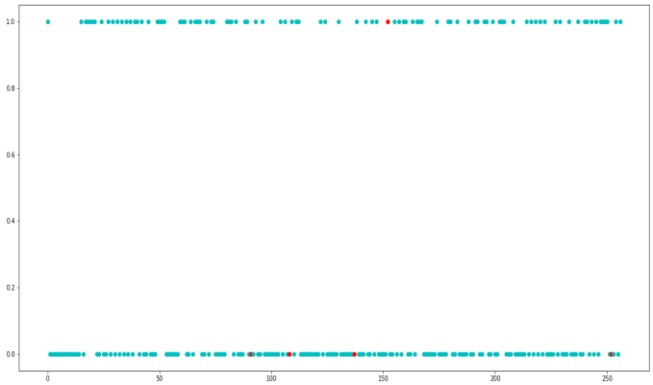

Figura 6. Valores de Previsão e Valores Verdadeiros

Nota: É necessário aumentar a imagem para visualizar os valores

O gráfico mostra que, para os dados correspondentes, o modelo falha 5 vezes, sendo 4 delas falsos negativos. Falsos negativos em cancros da mama apresentam maior risco do que falsos positivos.

Embora o método tenha uma precisão quase perfeita e alcance quase todos os valores, para a medicina, seria necessário ter menos falsos negativos para que o projeto fosse mais fiável.

Mesmo assim, posso afirmar que os resultados obtidos estiveram próximos da perfeição, falhando apenas 5 diagnósticos em 257. The graph shows that for the respective data the model fails 5 times, 4 of which represent false negatives.

7 Conclusões e Trabalho Futuro

O cancro da mama é uma neoplasia comum em mulheres em todo o mundo, e a deteção precoce pode melhorar significativamente o prognóstico e as chances de sobrevivência ao promover o tratamento clínico precoce dos pacientes.

Neste projeto, desenvolvi um programa de machine learning em Python para detetar cancro da mama a partir de um conjunto de dados. O objetivo deste trabalho é fornecer um prognóstico clínico precoce que possa eventualmente ajudar a salvar vidas apenas usando dados, Python e machine learning.

Neste artigo, falei sobre os algoritmos e técnicas que utilizei para criar esta inteligência artificial e torná-la útil na previsão do cancro.

Para concluir, se tivesse mais conhecimento sobre como melhorar os resultados no contexto do trabalho, talvez conseguisse reduzir os falsos negativos (mesmo que no pior caso induza mais falsos positivos) e, se possível, melhorar a precisão para atingir um valor próximo de 100%.

Referências

.

1. International Agency for Research on Cancer, Latest global cancer data: Cancer burden rises to 19.3 million new cases and 10.0 million cancer deaths in 2020, 15 December 2020 from https://www.iarc.who.int/wp-content/uploads/2020/12/pr292_E.pdf

2. Global Cancer Observatory: Cancer Today. Lyon, France: International Agency for Research on Cancer, from https://gco.iarc.fr/today/home

3. European Commission, Breast cancer burden in EU-27, European Cancer Information System, from https://ecis.jrc.ec.europa.eu/pdf/Breast_cancer_factsheet-Dec_2020.pdf

4. Cancer.net, Breast Cancer: Risk Factors and Prevention, July 2020, from https://www.cancer.net/cancer-types/breast-cancer/risk-factors-and-prevention

5. Polat, K., Günes S. (2006). Breast cancer diagnosis using least square support vector machine. Digital Signal Processing 17 (2007) 694–701. From https://reader.elsevier.com/reader/sd/pii/S1051200406001461?token=4628C65B06D2D675C621E5AA96F624445408EF01CB0535351515C66EF338B74E42B3DF976D94EECCECA90A18110CA3E8&originRegion=eu-west-1&originCreation=20211230190915

6. Paulin, F., Santhakumaran, A. (2011). Classification of Breast cancer by comparing Back propagation training algorithms. From http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.301.8837&rep=rep1&type=pdf

7. Dheeba, J., Singh, N., Selvi, S. (2014). Computer-aided detection of breast cancer on mammograms: A swarm intelligence optimized wavelet neural network approach. Journal of Biomedical Informatics 49 (2014) 45-52. From https://reader.elsevier.com/reader/sd/pii/S1532046414000124?token=C776CDAFB3458834333FD21CD7E3AA73B800BAE5A588D0B4C0A285684F1004432FBF812E53E1FCBE2F3AA02B5208C1B8&originRegion=eu-west-1&originCreation=20211230192914

8. Nahato, K., Harichandran, K., Arputharaj, k. (2015). Knowledge Mining from Clinical Datasets Using Rough Sets and Backpropagation Neural Network. From https://www.hindawi.com/journals/cmmm/2015/460189/

9. Logistic Regression - Detailed Overview, 15 March 2018, from https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc

10. Naïve Bayes Algorithm: Everything you need to know, from https://www.kdnuggets.com/2020/06/naive-bayes-algorithm-everything.html

11. Support Vector Machine, Introduction to Machine Learning Algorithms, 7 June 2018, from https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

12. *K*-Nearest Neighbors algorithm, from https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm

13. Decision Trees in Machine Learning, 17 May 2018, from https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052

14. Understanding Random Forest How the Algorithm Works and Why it Is So Effective, 12 June 2019, from https://towardsdatascience.com/understanding-random-forest-58381e0602d2

15. Neural Networks by IBM Cloud Education, 17 August 2020, from https://www.ibm.com/cloud/learn/neural-networks