Fundamentos de Inteligência Artificial

Docente: Drª Pétia Georgieva

Universidade de Aveiro – DFIS/DETI

2021/2022

Previsão de doença na tiroide

Marta Matos nº98479

Miguel Marques nº 98532

Resumo

A tiroide pode ser afetada por diversos problemas, sendo os principais, o bócio, as doenças autoimunes, o hipertiroidismo, o hipotiroidismo e o surgimento de nódulos. As doenças da tiroide são muito comuns, afetam mais 300 milhões de pessoas no mundo, sendo de um modo geral, mais comuns nas mulheres.

Uma doença na tiroide é o termo geral para a condição médica em que a tiroide está impedida de criar a quantidade certa de hormonas. Normalmente a tiroide produz hormonas q necessárias para o funcionamento normal do nosso corpo. Quando produz hormonas a menos sofremos de hipotireoidismo, quando produz a mais sofremos de hipertireoidismo. Estas hormonas controlam o nosso metabolismo, este processo transforma a comida que ingerimos em energia.

O nosso trabalho tem como objetivo prever se uma determinada pessoa sofre da doença na tiroide.

Introdução

Para alcançar o nosso objetivo, previsão de doença na tiroide, criamos um modelo de machine learning, neste modelo utilizamos um conjunto de dados [1] de várias pessoas que contêm várias *features* relevantes para a deteção da doença. No entanto não utilizamos todos os parâmetros, apenas alguns que foram selecionados.

Análise do conjunto de dados

O conjunto de dados contém 29 parâmetros, que são relevantes para a doença na tiroide, para além disso contém também uma *Label*, em que terá um valor binário, *0* para as pessoas sem a doença e *1* para as pessoas que a têm.

Descrição de features:

age – idade da pessoa

sex – gênero da pessoa

on_thyroxine – se a pessoa está a tomar medicação para tiroxina

query_on_thyroxine – se foi medido a sua tiroxina

on_antithyroid_medication – se a pessoa está a tomar medicação para a tiroide

sick – se a pessoa está doente

pregnant – se a pessoa está grávida

thyroid_surgery – se a pessoa fez uma cirurgia

I131_treatment – se a pessoa fez um tratamento médico para tratar uma tiroide hiperativa

query_hypothyroid – se a pessoa sofre de hipotireoidismo

query_hyperthyroid – se a pessoa sofre de hipertireoidismo

goitre – aumento do volume da glândula tiroide

lithium – se a pessoa está a usar lítio(medicação), esta leva ao desenvolvimento de hipotireoidismo, hipertireoidismo e bócio

tumor – se a pessoa já teve um tumor

hypopituitary – se a pessoa é afetada por uma doença endócrina, esta é caraterizada pela diminuição da secreção de um ou mais das oito hormonas produzidas pela glândula pituitária

psych – se a pessoa sofre de alterações rápidas nas hormonas da tiroide, estas podem perturbar as emoções da mesma

T3 – quantidade de triiodotironina que a pessoa contém, esta hormona da tiroide sintetizada primeiramente nos tecidos periféricos a partir da tiroxina, esta também é secretada em pequenas quantidades pela glândula tiroide, as unidades da T3 são (ng/dL)

T3_measured – se foi medido o nível da hormona T3

TT4 – é o total de tiroxina da pessoa, esta é uma hormona produzida na tiroide e posteriormente é direcionada para a corrente sanguínea, as unidades de TT4 são (µg/dL)

TT4_measured – se foi medido o nível total de tiroxina

T4U – é a quantidade de tiroxina que a pessoa utiliza, as unidades de T4U são (µg/dL)

T4U_measured – se foi medido o nível da tiroxina utilizada

FTI – índice de tiroxina livre na pessoa, as unidades de FTI são (µg/dL)

FTI_measured – se foi medido o nível de tiroxina livre na pessoa

TSH – quandidade da hormona tireoestimulante que a pessoa contém, a hormona produzida pela hipófise e tem como objetivo estimular a produção das hormonas T3 e T4 por parte da tiroide, as unidades da TSH são (µUI/mL)

TSH_measured – se foi medido o nível da hormona TSH

TBG – a globulina ligadora de tiroxina é uma proteína responsável pelo transporte das hormonas na tiroide, esta proteína é a que possui mais afinidade com T3 e T4

TBG_measured – se foi medido o nível da proteína TBG

referral_source – método de deteção da doença da tiroide

class – é a label da base de dados, e classifica as pessoas que tem a doença e as que não a tem

Pré Processamento de dados

As *features* do nosso conjunto de dados estavam em *strings* e em *float* pelo que tivemos de transformar as *strings* em binário, e normalizar os valores de float, as *features* que foram normalizadas adicionámos "norm" no início do seu nome.

Como o nosso conjunto de dados têm várias *features*, nós fizemos a seleção de algumas utilizando uma matriz de correlação. Na qual filtramos a coluna da *Class* que era a que nos interessava.

features	Coeficiente de correlação
sex	0.031
on_thyroxine	0.048
on_antithyroid_medication	0.033
sick	0.096
pregnant	0.037
thyroid_surgery	0.033
I131_treatment	0.025
query_hypothyroid	0.087
query_hyperthyroid	0.035
lithium	0.002
goitre	0.009
tumor	0.025
hypopituitary	0.066
psych	0.043
Class	1.000
normTSH	0.008
normT3	0.397
normTT4	0.127
normT4U	0.243
normFTI	0.022
normAge	0.165

Tabela 1 Matriz de correlação da feature Class

Baseado na "Tabela 1" selecionámos as *features* com um coeficiente de correlação superior a 0.10. Ficando com os seguintes parâmetros: normT3, normTT4, normT4U, normAge, Class.

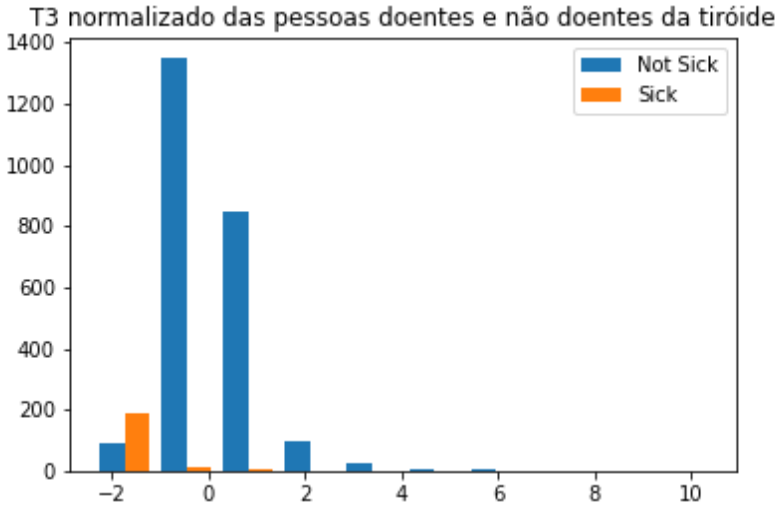

Gráfico 1 Comparação da distribuição de T3 normalizado entre as pessoas doentes e não doentes da tiroide

Através do gráfico 1 podemos verificar que as pessoas com doença na tiroide têm uma quantidade menor de T3, os a hormona T3 das pessoas que não têm doença segue uma distribuição normal

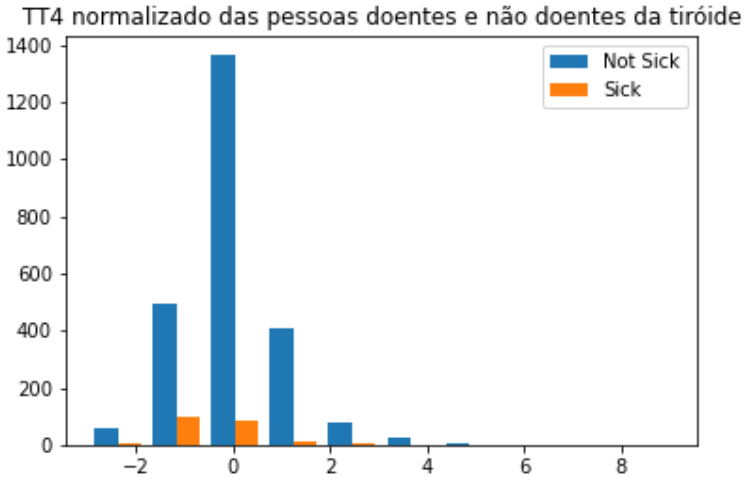

Gráfico 2 Comparação da distribuição de TT4 normalizado entre as pessoas doentes e não doentes da tiroide

Analisando o gráfico 2, podemos confirmar que a distribuição do total da hormona T4 segue uma distribuição normal em ambos os casos.

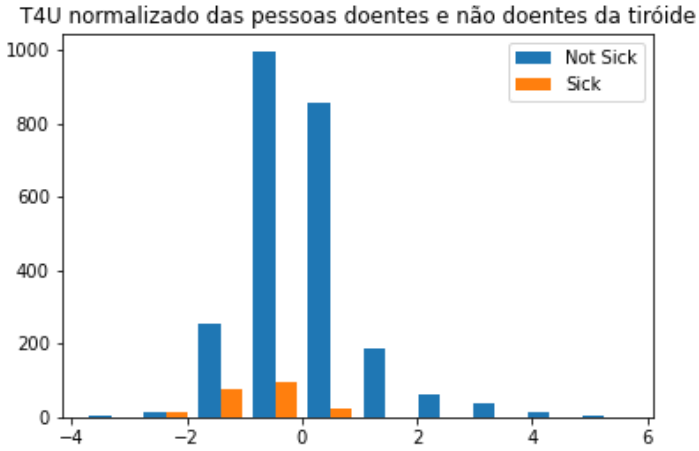

Gráfico 3Comparação da distribuição de T4U normalizado entre as pessoas doentes e não doentes da tiroide

Analisando o gráfico 3, podemos confirmar que a distribuição da utilização da hormona T4 segue uma distribuição normal em ambos os casos, no entanto os que sofrem da doença tem valores relativamente mais baixos.

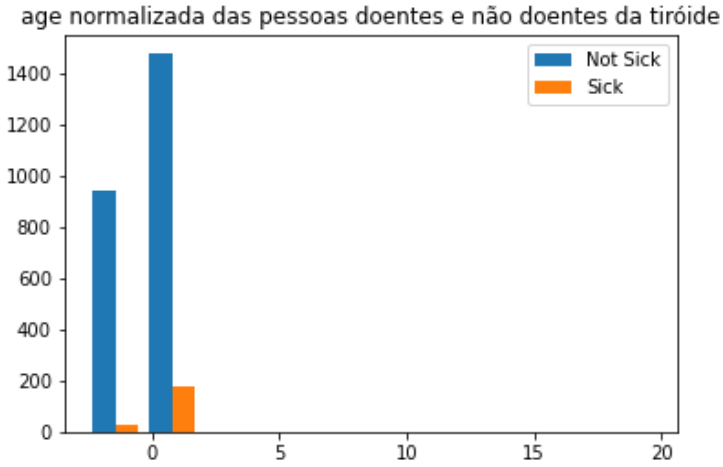

Gráfico 4 Comparação da distribuição de age normalizada entre as pessoas doentes e não doentes da tiroide

Após observarmos o gráfico 4, podemos concluir que maior parte das pessoas com doença na tiroide tem uma idade elevada.

Após esta seleção fazemos a contagem do número de pessoas com doença na tiroide e as que não a têm.

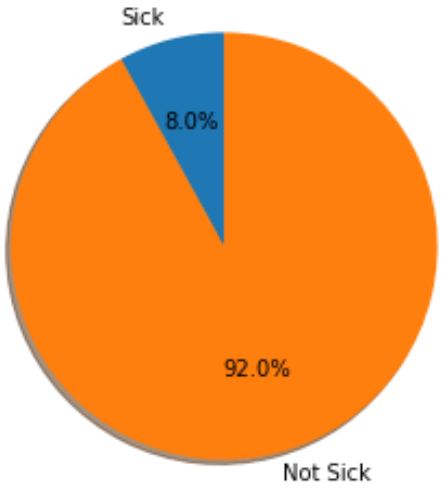

Gráfico 5 Percentagem de pessoas com doença na tiroide, assim como, as que não a tem

Como podemos observar existe uma desigualdade enorme entre o número de pessoas com a doença em comparação com as que não a tem. Por estes motivos iremos recorrer a um processo *under-sampling*, este processo consiste em igualar a *Class* que está em minoria, neste caso os que têm doença na tiroide com a *Class* que está em maioria, os que não tem doença. Após o este processo obtemos:

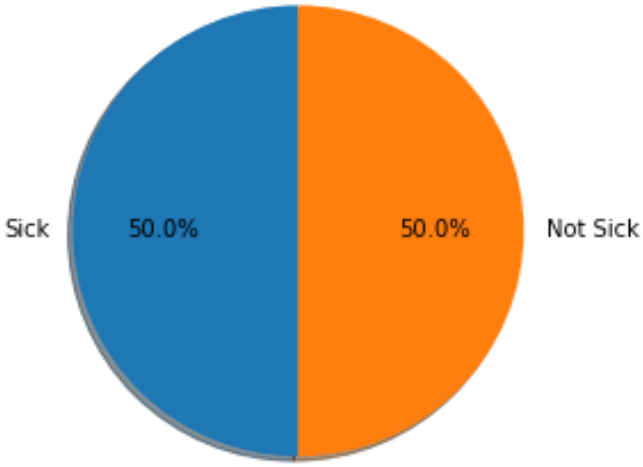

Gráfico 6 Percentagem de pessoas com doença na tiroide, assim como, as que não a tem, após o processo under-sampling

Treino, Validação e Teste

Tendo concluído o pré-processamento dos dados, fizemos a separação em conjuntos de treino e teste, isto foi realizado com o conjunto total de dados assim como com o conjunto de dados *under-sampled*. Inicialmente calculamos o *hyper-parameter* com o conjunto de dados *under-sampled* utilizando o *K-fold Cross Validation*, após isto fazemos uma *Logistic Regression* utilizando o hyper-parameter, depois treinamos a *Logistic Regression* e por fim testamos os dados. Por fim é possível calcular a matriz confusão, através desta podemos determinar a *Precision, Accuracy e Recall* do modelo.

Accuracy = $\frac{(TP+TN)}{Total}$

Precision = $\frac{TP}{(TP+FP)}$

Recall = $\frac{TP}{(TP+FN)}$

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

O nosso objetivo é prever as pessoas que tem doença na tiroide pelo que apenas nos interessa o *Recall.*

Para a *Logistic Regression* treinada e testada com os dados *under-sampled.*

52	7
5	64

Tabela 2 Matriz confusão para a Logistic Regression treinada e testada com os dados under-sampled

Accuracy = $\frac{52+64}{52+7+5+64} = 60{,}68$

Precision = $\frac{52}{52+7} = 0{,}90$

Recall = $\frac{52}{52+5} = 0{,}93$

F1 Score = $\frac{2*(0{,}93*0{,}90)}{(0{,}93\ 0{,}90)} = 0{,}91$

Para a regressão logística treinada com os dados *under-sampled* e testada com o conjunto total de dados de teste.

647	84
8	54

Tabela 3 Matriz Confusão para a Logistic Regression treinada com os dados under-sampled e testada com o conjunto total de dados de teste

$$Precision = \frac{647}{(647+84)} = 0{,}39$$

$$Recall = \frac{52}{52+5} = 0{,}87$$

$$\text{F1 Score} = \frac{2*(0{,}93*0{,}90)}{(0{,}93\ 0{,}90)} = 0{,}54$$

Por fim calculámos a curva de *Receiver Operating Characteristic (ROC), assim como cálculamos a área* sob a curva de *ROC*, denominada de AUC. Como os dados que utilizámos foram *under-sampled* devemos executar o modelo múltiplas vezes e calcular a curva ROC e a sua área para a garantir que o modelo funciona para vários casos.

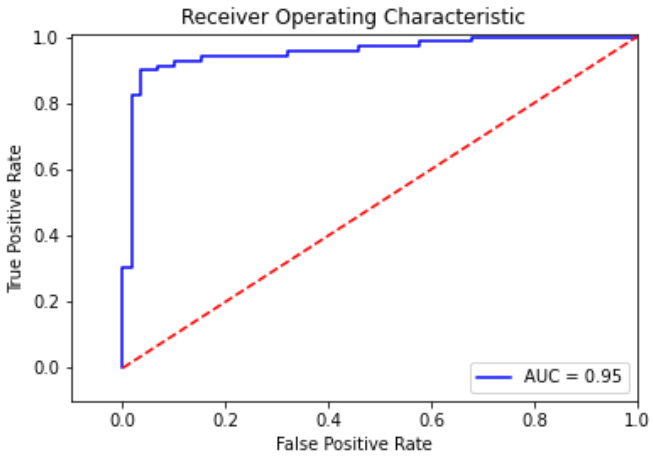

Gráfico 7 Curva ROC na primeira vez que executamos o modelo

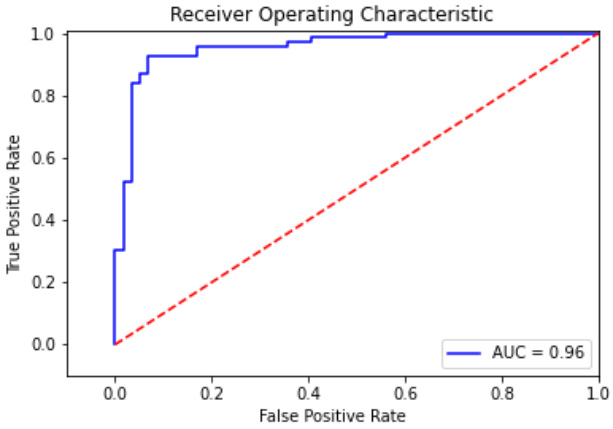

Gráfico 8 Curva ROC na segunda vez que executamos o modelo

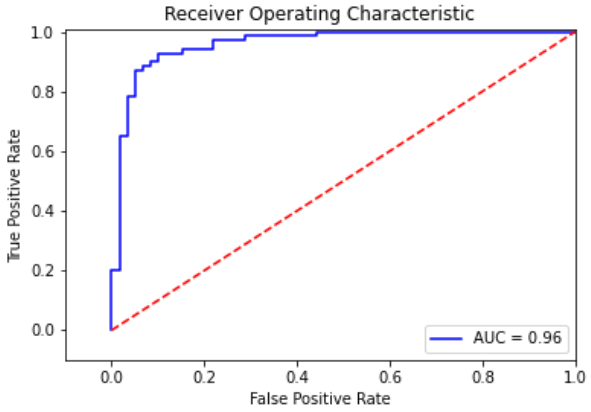

Gráfico 9 Curva ROC na terceira vez que executamos o modelo

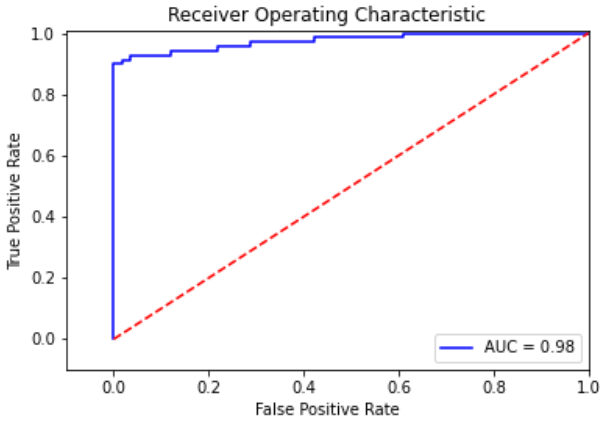

Gráfico 10 Curva ROC na quarta vez que executamos o modelo

Como podemos observar pelas quatro execuções do modelo, obtivemos os seguintes valores para AUC: 0.95, 0.96, 0.96, 0.98 o que nos leva a concluir que o modelo é relativamente bom, AUC > 0.90, mas não é perfeito uma vez que o AUC é sempre inferior a 1.

Conclusão

O modelo que obtivemos para a previsão de uma doença na tiroide, apresenta um valor de 0,93 e 0,87 no *Recall* da regressão linear treinada e testada com os dados *under-sampled,* da regressão linear treinada com os dados *under-sampled* e testada com o conjunto total de dados de teste respetivamente, o que foi satisfatório. Os valores obtidos para AUC nas quatros vezes que executamos o modelo não foram perfeitos, mas foram bons. Pelo que podemos afirmar que o objetivo deste trabalho foi alcançado com sucesso. O nosso modelo pode ser melhorado adicionando mais *features* no entanto conforme adicionamos será necessário um poder computacional mais elevando para executar o modelo.
Comparando o nosso modelo com [2], o modelo de Bertie obteve um *Recall* igual a 91,3% o que é inferior ao primeiro *Recall* calculado no nosso modelo, e superior ao segundo, ou seja, utilizando o primeiro *Recall podemos afirmar que o nosso modelo é melhor que o do Bernie.*

Referências

[1] Thyroid disease records collected and supplied by the Garavan Institute and J. Ross Quinlan, New South Wales Institute, Sydney, Australia. 1987.

[2] Bertie, RF classifier using different feature selections