Previsão de Doença na Tiroide

Fundamentos de Inteligência Artificial

–

Universidade de Aveiro

2021/2022

Marta Matos (NMec 98479)
Licenciatura em Engenharia Computacional
Departamento de Física
Aveiro, Portugal
martacarolinafm@ua.pt
Percentagem de trabalho: 50%

Miguel Marques (NMec 98532)
Licenciatura em Engenharia Computacional
Departamento de Física
Aveiro, Portugal
miguel.rosas@ua.pt
Percentagem de trabalho: 50%

***Abstract*—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 necessárias para o funcionamento normal do nosso corpo. Quando produz hormonas a menos sofremos de hipotiroidismo, quando produz a mais sofremos de hipertiroidismo. 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, para isso, pretendemos comparar vários modelos de classificação incluindo Hidden Markov Model e Bayesian Network.

***Index Terms*—Previsão de Doença na Tiróide, Machine Learning, Hidden Markov Model, Bayesian Network**

I. Introdução

O desenvolvimento de novas tecnologias e métodos de inteligência artificial têm tido um crescimento acentuado, pelo que tem tornado possível e mais prática a capacidade de determinação da doença na tiroide.

Este projeto foi realizado no âmbito da unidade curricular "Fundamentos de Inteligência Artificial" do Departamento de Electrónica, Telecomunicações e Informática (DETI) na Universidade de Aveiro, lecionada pelos professores Pétia Georgieva e Engénio Rocha.

Para alcançar o nosso objetivo, previsão de doença na tiroide, criamos um modelo de machine learning, no qual 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.

Este projeto encontra-se dividido em quatro partes: a primeira, secção II, na qual fizemos a análise do conjunto de dados, seguida da secção III com o pré-processamento dos mesmos. Na secção IV descrevemos os modelos utilizados e por fim, na secção V, apresentamos uma visão geral de todo o trabalho realizado.

II. 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.

A. Descrição de features

- `age` – idade da pessoa
- `sex` – género da pessoa
- `on_thyroxine` – se a pessoa está a tomar medicação para tiroxina
- `query_on_thyroxine` – se foi medida 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 hipotiroidismo
- `query_hyperthyroid` – se a pessoa sofre de hipertiroidismo
- `goitre` – aumento do volume da glândula tiroide
- `lithium` – se a pessoa está a usar lítio(medicação), esta leva ao desenvolvimento de hipotiroidismo, hipertiroidismo 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 (UI/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

III. 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, fizemos a seleção de algumas utilizando uma matriz de correlação, na qual filtramos a coluna da Class que era a que nos interessava.

TABLE I
Tabela de coeficientes de relação

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

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

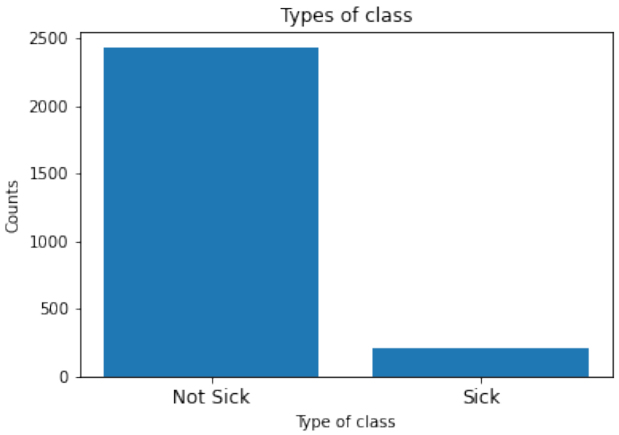

Fig. 1. Gráfico da percentagem de pessoas com doença na tiroide, assim como, as que não a tem

Como podemos observar existe uma desigualdade enorme entre o número de pessoas com a doença comparativamente com as que não a tem.

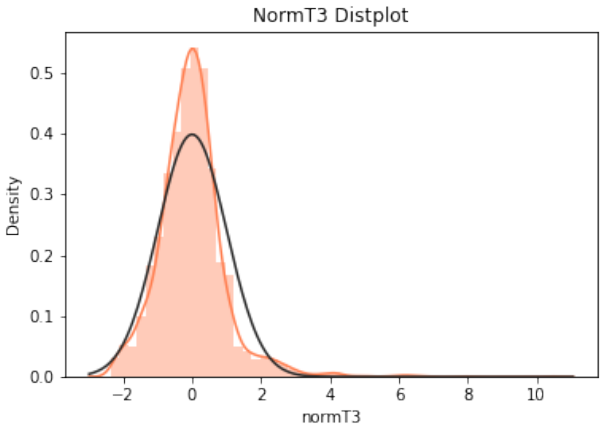

Fig. 2. Gráfico de comparação da distribuição de T3 normalizado das pessoas com doença e pessoas sem doença.

Podemos observar que os dados foram corretamente normalizados.

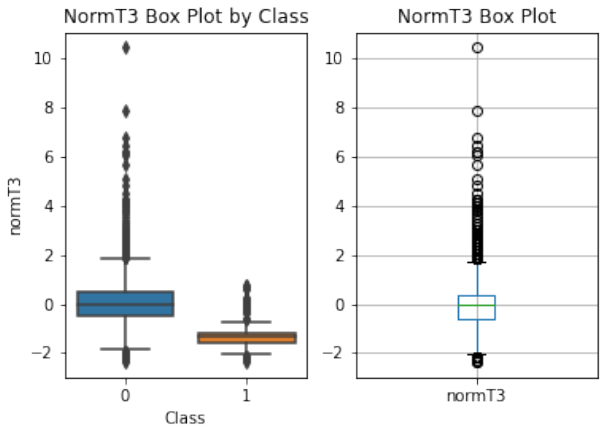

Fig. 3. Box Plot da feature normT3 por valor da Class, e da normT3

Podemos analisar que os valores da Class 0 entre os quartis 0.25 e 0.75 são superiores aos da Class 1.

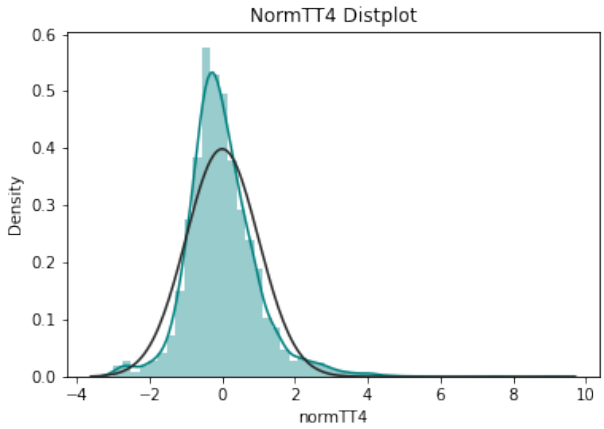

Fig. 4. Gráfico de comparação da distribuição de TT4 normalizado das as pessoas doentes e não doentes da tiroide.

Em que mais uma vez vizualizamos que a normalização foi efetuada com sucesso.

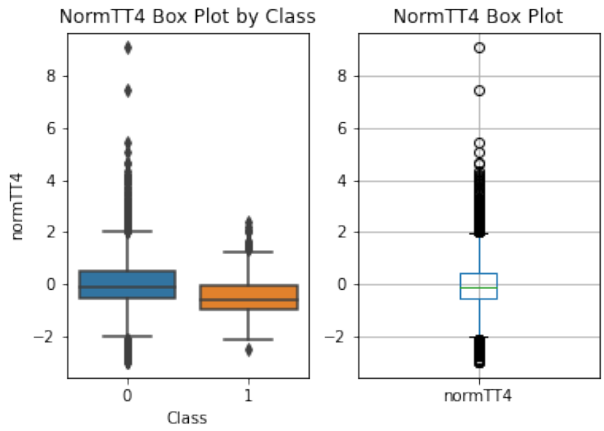

Fig. 5. Box Plot da feature normTT4 por valor da Class, e da normTT4

Podemos analisar que os valores da Class 0 entre os quartis 0.50 e 0.75 são superiores aos da Class 1.

IV. Descrição dos Modelos utilizados

Inicialmente utilizamos o modelo Logistic Regression e Decision Tree Classifier, para classificar para além destes no estudo de previsão de doença na tiroide, num determinado indivíduo, foi então implementado para comparação o modelo de classificação Hidden Markov Model (HMM) e o Bayesian Network (BN).

A. *Logistic Regression*

1) Descrição:

A regressão logistica estima a probabilidade de um evento occorrer, baseados no data set de variáveis independentes. O resultado da regressão logistica é uma probabilidade que é arredondada para 0 ou 1.

2) Resultados:

	precision	recall	f1-score	support
0	0.956291	0.987688	0.971736	731.000000
1	0.763158	0.467742	0.580000	62.000000
accuracy	0.947037	0.947037	0.947037	0.947037
macro avg	0.859725	0.727715	0.775868	793.000000
weighted avg	0.941191	0.947037	0.941109	793.000000

Fig. 6. Tabela sobre os resultados da Logistic Regression

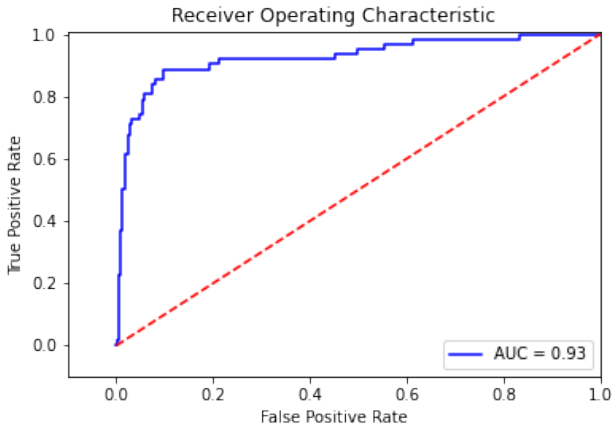

Fig. 7. Gráfico da curva ROC

Calculámos a curva de Receiver Operating Characteristic (ROC), assim como cálculamos a área sob a curva de ROC, denominada de AUC. Os resultados foram satisfatórios para a logistic regression, uma vez que a AUC = 0.93.

B. *Decision Tree Classifier*

1) Descrição:

A tree decision é uma ferramenta de suporte de um modelo do tipo árvore de decisões e as suas possiveis consequências, inclui os resultados dos eventos por chance, os custos e a utilidade.

2) Resultados:

	precision	recall	f1-score	support
0	0.982167	0.979480	0.980822	731.000000
1	0.765625	0.790323	0.777778	62.000000
accuracy	0.964691	0.964691	0.964691	0.964691
macro avg	0.873896	0.884901	0.879300	793.000000
weighted avg	0.965237	0.964691	0.964947	793.000000

Fig. 8. Tabela sobre os resultados do Tree Decision Classifier

No *decision tree classifier*, obtivemos resultados melhores a nível da precision, f1-score e accuracy.

C. *Hidden Markov Model*

1) Descrição:

Hidden Markov Model (HMM) ou, traduzindo, Modelo oculto de Markov, trata-se de um modelo estatístico no qual o sistema modelado se assume com um processo de Markov com parâmetros desconhecidos, cujo objetivo é a determinação dos parâmetros ocultos através dos parâmetros observáveis, e cujos parâmetros obtidos podem ser usados na realização de novas análises.

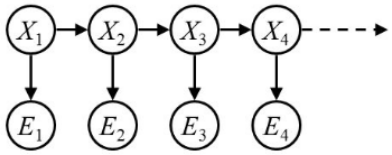

Fig. 9. Esquema de HMM

Um HMM consiste em:

- Um conjunto de estados X (geralmente assumido como finito).
- Uma distribuição de estado inicial $P(X_1 = x), x \in X$. Este anotado o nó superior esquerdo no modelo gráfico (Figura 7).
- Probabilidades de transição de estado: $P(X_{t+1} = x'|X_t = x), x, x \in X$. Estes anotam os arcos à direita no modelo gráfico (Figura 7).
- Um conjunto de observações E (muitas vezes assumidas como finitas).
- Probabilidades de observação de emissões $P(E_t = e|X_t = x), x \in X, e \in E$. Estes anotam os arcos em baixo acima (Figura 7).

Esse tipo de modelo é conhecido em variadas áreas de reconhecimento de padrões temporais como a fala, os gestos, a escrita, e a bioinformática.

2) Implementação:

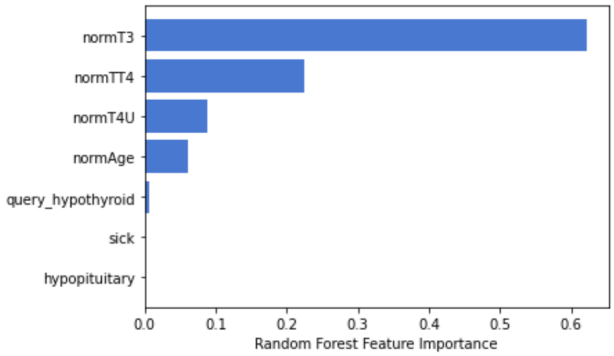

Fig. 10. Gráfico de determinação de features mais importantes

Na implementação do modelo tivemos que reduzir o nosso data set original, para tal, escolhemos as duas características mais importantes de forma a obter melhores resultados. Verificando o gráfico da Figura X concluímos que as features eleitas são normT3 e normTT4. Categorizamos as features baseando-nos em percentis, seguidamente cálculamos o número de transições e as probabilidades de emissões, com isto construímos o modelo e testamos-o mais tarde. Na construção do modelo utilizamos o data set, no teste usamos um grupo de teste obtido através do data set original.

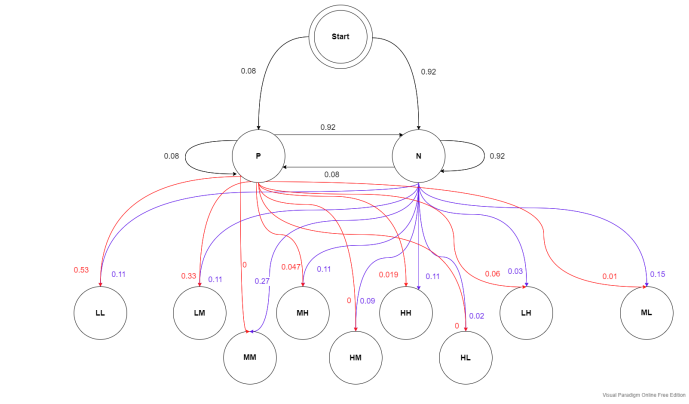

Fig. 11. Gráfico do Modelo Markov

No nosso gráfico do Modelo Markov, podemos observar o custo para os states P e N, que representam as pessoas com doença na tiróide, e as pessoas que não sofrem da doença. Através dos states podemos vizualizar o custo das observations e o custo de mudança de state como de permanência.
As observações são estão identificadas de maneira em que a primeira letra categoriza a feature normT3 e a segunda letra categoriza a feature normTT4. As categorias para estas duas features são três, L sigifica Low, M significa Medium e H significa High. Estas categorias foram classificadas baseados em percentis das próprias features.

3) Validação e Resultados:

Para validar os resultados obtidos após ter aplicado o Modelo Markov ao *test data set*, comparamos a *classicação prevista* com a *classificação tabelada*. Contamos o número de classificações corretas e devidimos pelo número total de entradas no *test data set*, através deste cálculo obtivemos uma accuracy 92,18%, o que é bom. No entanto o nosso data set não é balanceado o que pode afetar os resultados, nomeadamente a accuracy.

*D. **Bayesian Network***

1) Descrição:

Bayesian Network (BN) ou Rede Bayesiana, consiste num grafo que representa uma ou mais relações de probabilidade condicional, como por exemplo, a ocorrência de certas variáveis depende do estado de outra, facilitando a predição e "abdução" dos sistemas de inteligência artificial. BNs são modelos gráficos de raciocínio baseado em incerteza, nos quais os nós representam as variáveis, que podem ser discretas ou contínuas, e os arcos representam as conexões diretas entre eles.

2) Implementação:

A representação de uma rede Bayesiana só é considerada correcta no domínio se cada nó for condicionalmente independente dos seus antecessores, considerando os seus pais, isto é, cada nó é condicionalmente independente dos seus pares e independente dos seus predecessores na ordenação, dado o seu nó pai.

Uma BN é representável por uma solução fatorial de uma distribuição conjunta que se pode dividir em duas partes, a estrutura e os parâmetros. A sua estrutura assemelha-se à de um gráfico acíclico, sem ciclos de nós, que representa as dependências e independências condicionais entre as diferentes variáveis associadas aos nós dos dados fornecidos. Os parâmetros são produto das distribuições de probabilidade condicional associadas a cada nó.

Na aplicação da Bayesian Network ao nosso data set, fizemos um *under-sampling* para balancear o data set, após isto separamos o data set em dois, um para o treino outro para o teste. Classificamos categoricamente o data set de treino e teste. Tendo isto feito calculamos a probabilidade para cada categoria das features e fizemos também o mesmo cálculo mas desta vez em função da Classe, para além disso calculamos as probabilidades das combinações da features. No final disto testamos os dados, tendo em conta as probabilidades de cada categoria das features em função da Classe, escolhemos a Classe que tinha maior probabilidade.

3) Validação e Resultados:

Para a validação dos dados, após o teste comparamos os resultados obtidos da classe, com os valores originais. Para calcular a accuracy contamos a quantidade de classes previstas iguais às originais e dividimos pelo número de entradas no data set de treino. Como nós aplicamos *under-sampling* executamos o código múltiplas vezes, uma vez que os grupos de teste e treino são escolhidos aleatoriamente. Após múltiplas execuções verificamos que os valores obtidos para a accuracy rondava os 70%, que é um resultado satisfatório.

V. Conclusão

O modelo que obtivemos para a previsão de uma doença na tiróide, para o *Hidden Markov Model*, obtivemos uma accuracy de 92,18%, o que é bom, no entanto, como foi referido anteriormente o data set não está balanceado o que pode after o resultado. No caso do *Bayesian Network*, aplicámos *under-sampling* ao data set, e após várias execuções do modelo obtivemos uma accuracy que rondava os 80%, o que é satisfatório.

References

[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.