Universidade Federal de São João del-Rei

Ciência da Computação

Análise e Classificação de Doenças da Tireoide com XGBoost

22 de setembro de 2025

Relatório de Mineração de Dados

Kariny Abrahão (212050013)

Professor:
Leonardo Rocha

Sumário

1 Introdução

A glândula tireoide desempenha um papel fundamental na regulação do metabolismo humano por meio da produção dos hormônios triiodotironina (T3) e tiroxina (T4). Alterações no funcionamento dessa glândula podem gerar duas condições clínicas principais: **hipotireoidismo** e **hipertireoidismo**.

O **hipotireoidismo** ocorre quando há uma produção insuficiente de hormônios tireoidianos, o que pode levar a sintomas como fadiga, ganho de peso, intolerância ao frio e depressão. Já o **hipertireoidismo** é caracterizado pelo excesso de hormônios tireoidianos, causando perda de peso não intencional, ansiedade, taquicardia e intolerância ao calor. Ambas as condições, quando não diagnosticadas ou tratadas adequadamente, podem comprometer significativamente a qualidade de vida dos pacientes e até gerar complicações mais graves.

Na prática clínica, o diagnóstico dessas disfunções é realizado principalmente pela análise de exames laboratoriais. O **TSH (hormônio estimulador da tireoide)** é considerado o marcador mais sensível: níveis elevados de TSH sugerem **hipotireoidismo**, enquanto níveis suprimidos (muito baixos) indicam **hipertireoidismo**. Além disso, a dosagem dos hormônios tireoidianos T3 e T4 auxilia na confirmação do quadro: no hipotireoidismo observa-se T4 reduzido, enquanto no hipertireoidismo os níveis de T3 e T4 costumam estar aumentados. Outros índices derivados, como o **FTI (Free Thyroxine Index)** e o **T4U (T4 uptake)**, também são utilizados para refinar a interpretação.

Apesar de bem estabelecidos na prática médica, esses exames nem sempre apresentam cortes claros entre as diferentes condições, havendo sobreposição de valores entre indivíduos saudáveis e doentes. Esse cenário torna o diagnóstico desafiador em alguns casos, principalmente quando existem valores limítrofes ou resultados laboratoriais inconsistentes.

Nesse contexto, técnicas de **mineração de dados e aprendizado de máquina** vêm sendo exploradas como ferramentas de apoio ao diagnóstico de doenças da tireoide. Essas técnicas permitem lidar com grandes conjuntos de dados, aplicar estratégias de pré-processamento para tratar inconsistências e treinar modelos de classificação capazes de distinguir entre indivíduos saudáveis, com hipotireoidismo ou com hipertireoidismo.

O presente trabalho tem como objetivo realizar um estudo de caso utilizando a base de dados *Thyroid Disease Data*, disponível no Kaggle [1] para responder:

Como técnicas de aprendizado de máquina, em especial o XGBoost, podem auxiliar no diagnóstico automatizado de doenças da tireoide (hipotireoidismo e hipertireoidismo) a partir de exames laboratoriais?

---

2 Atividades Práticas

Durante o desenvolvimento deste trabalho prático, foram realizadas as seguintes atividades:

- Pesquisa bibliográfica sobre doenças da tireoide (hipotireoidismo e hipertireoidismo) e revisão de artigos que utilizam aprendizado de máquina aplicado ao diagnóstico médico.
- Seleção e análise inicial dos dados.
- Pré-processamento dos dados, análise exploratória dos dados, tratamento de valores, padronização e balanceamento das classes utilizando *SMOTE*.
- Implementação de modelos de classificação em Python, com ênfase no algoritmo XGBoost.
- Avaliação dos modelos utilizando métricas como acurácia, precisão, recall e f1-score.
- Geração de gráficos e visualizações para análise das importâncias das variáveis no modelo.
- Redação do relatório em LaTeX, contendo introdução teórica, metodologia, resultados obtidos e considerações finais.

2.1 Atividade 1 - Pesquisa bibliográfica

Para fundamentar o pré-processamento e as técnicas de classificação aplicadas neste trabalho, foram utilizadas como referência diversas fontes, entre notebooks e artigos científicos. Em particular:

- Notebooks do Kaggle, como "XGBoost Multi-Class Classification" de EmmanuelFwerr, "Thyroid Disease Detection" de iamArslanKhalid, e "Thyroid Classification" de ZiadAbdelaziz, que fornecem exemplos práticos de análise exploratória de dados (EDA), tratamento de valores ausentes, encoding e avaliação com métricas variadas.
- O artigo "*Enhanced Diagnosis of Thyroid Diseases Through Advanced Machine Learning Methodologies*" (Oture, Iqbal & Wang, 2025), que compara várias técnicas de ML e DL aplicadas ao mesmo tipo de dados, ressaltando uso de oversampling/undersampling e identificando TSH como biomarcador importante. [2]

Essas referências ajudaram a definir escolhas no meu pipeline, como quais variáveis manter, quais métricas observar, como tratar valores ausentes e como balancear classes para evitar viés no modelo.

2.2 Atividade 2 - Seleção e Análise Exploratória Inicial dos Dados

A segunda atividade teve como objetivo compreender melhor o conjunto de dados *Thyroid Disease Data*, disponível no Kaggle [1], que contém informações clínicas e laboratoriais de pacientes, incluindo variáveis contínuas e categóricas. O dataset possui 9172 registros, quantidade considerada suficiente para a condução de nosso estudo, permitindo a exploração de padrões relevantes e o treinamento de modelos de aprendizado de máquina com validade estatística.

O foco dessa etapa foi detalhar cada uma das features, entendendo o significado de seus valores e como elas poderiam impactar os diagnósticos médicos.

A Tabela 1 apresenta a descrição das colunas do dataset e os tipos de valores que elas contêm. A coluna **target** contém os diagnósticos médicos dos pacientes, que estão detalhados na Tabela 2.

Neste estágio, com a ajuda da análise detalhada das features, avaliamos o conjunto de dados e concluímos que ele é suficiente e adequado para o objetivo deste trabalho, que consiste no diagnóstico automatizado de hipotireoidismo e hipertireoidismo.

Tabela 1: Descrição das variáveis do conjunto de dados *Thyroid Disease Data*.

Variável	**Descrição**
age	Idade do paciente (inteiro)
sex	Sexo com o qual o paciente se identifica (string)
on_thyroxine	Se o paciente está em uso de tiroxina (booleano)
query_on_thyroxine	Se o paciente acredita estar em uso de tiroxina (booleano)
on_antithyroid_meds	Se o paciente está em uso de medicamentos antitireoidianos (booleano)
sick	Se o paciente está doente (booleano)
pregnant	Se o paciente está grávida (booleano)
thyroid_surgery	Se o paciente já passou por cirurgia na tireoide (booleano)
I131_treatment	Se o paciente está em tratamento com I131 (booleano)
query_hypothyroid	Se o paciente acredita ter hipotireoidismo (booleano)
query_hyperthyroid	Se o paciente acredita ter hipertireoidismo (booleano)
lithium	Se o paciente faz uso de lítio (booleano)
goitre	Se o paciente apresenta bócio (booleano)
tumor	Se o paciente possui tumor (booleano)
hypopituitary	Se o paciente apresenta hipopituitarismo (booleano)
psych	Se o paciente apresenta condição psiquiátrica (booleano)
TSH_measured	Se o TSH foi medido no sangue (booleano)
TSH	Nível de TSH no sangue a partir de exame laboratorial (float)
T3_measured	Se o T3 foi medido no sangue (booleano)
T3	Nível de T3 no sangue a partir de exame laboratorial (float)
TT4_measured	Se o TT4 foi medido no sangue (booleano)
TT4	Nível de TT4 no sangue a partir de exame laboratorial (float)
T4U_measured	Se o T4U foi medido no sangue (booleano)
T4U	Nível de T4U no sangue a partir de exame laboratorial (float)
FTI_measured	Se o FTI foi medido no sangue (booleano)
FTI	Nível de FTI no sangue a partir de exame laboratorial (float)
TBG_measured	Se o TBG foi medido no sangue (booleano)
TBG	Nível de TBG no sangue a partir de exame laboratorial (float)
referral_source	Fonte de encaminhamento do paciente (string)
target	Diagnóstico médico (string)
patient_id	Identificador único do paciente (string)

Tabela 2: Descrição dos códigos de diagnóstico presentes na coluna *target*.

Letra	Diagnóstico
Condições de hipertireoidismo	
A	Hipertireoidismo
B	T3 tóxico
C	Bócio tóxico
D	Tóxico secundário
Condições de hipotireoidismo	
E	Hipotireoidismo
F	Hipotireoidismo primário
G	Hipotireoidismo compensado
H	Hipotireoidismo secundário
Proteína de ligação	
I	Proteína de ligação aumentada
J	Proteína de ligação diminuída
Saúde geral	
K	Doença não tireoidiana concomitante
Terapia de reposição	
L	Consistente com terapia de reposição
M	Sub-reposição
N	Super-reposição
Tratamento antitireoidiano	
O	Medicamentos antitireoidianos
P	Tratamento com I131
Q	Cirurgia
Diversos	
R	Resultados de exames discordantes
S	TBG elevado
T	Hormônios tireoidianos elevados

2.3 Atividade 3 - Pré-processamento, Balanceamento dos Dados e Análise Exploratória

Após a análise inicial, iniciou-se o pré-processamento dos dados, etapa fundamental para garantir que o modelo de aprendizado de máquina pudesse ser treinado de forma eficaz. As ações realizadas incluíram:

2.3.1 Remoção de colunas redundantes e irrelevantes

O primeiro passo no pré-processamento foi analisar a relevância de cada coluna do dataset. Colunas do tipo `*_measured` foram removidas, pois indicavam apenas se um exame havia sido realizado e não forneciam informação adicional relevante para o modelo. A coluna `patient_id`, por se tratar de um identificador único, também foi descartada por não possuir utilidade preditiva. Da mesma forma, a coluna `referral_source` foi removida, por não agregar valor significativo na previsão do diagnóstico. Por fim, a coluna `TBG` foi excluída, uma vez que apresentava uma quantidade extremamente elevada de valores nulos (8823 de 9172, aproximadamente 96%) e não é considerada fundamental para os diagnósticos de hipotireoidismo ou hipertireoidismo.

Essas remoções simplificaram o dataset, reduziram a complexidade do modelo e eliminaram colunas que poderiam introduzir ruído ou viés desnecessário.

2.3.2 Limpeza e padronização da variável target

Posteriormente, iniciamos a exploração e limpeza da variável **target**, que representa o alvo do modelo. Primeiramente, foram removidos espaços em branco e todas as letras foram padronizadas para maiúsculas, evitando erros ou perda de informação.

Em seguida, filtramos apenas os diagnósticos relevantes para o estudo: as letras que representam o hipertireoidismo (A, B, C, D), o hipotireoidismo (E, F, G, H), e os casos negativos, representados pelo caractere `"-"`. Todas as demais linhas contendo outros diagnósticos foram removidas, a fim de evitar ruídos e possíveis vieses no modelo.

Por fim, a coluna **target** foi mapeada numericamente: 0 para os casos negativos, 1 para o hipotireoidismo e 2 para o hipertireoidismo, tornando-a adequada para o treinamento do modelo de classificação.

3 Desafios

4 Resultados

5 Conclusão

6 Reflection

7 Acknowledgements

Referências

[1] Emmanuel Fwerr. *Thyroid Disease Data*. https://www.kaggle.com/datasets/emmanuelfwerr/thyroid-disease-data. Acessado em: 21 set. 2025. 2023.

[2] Osasere Oture, Muhammad Zahid Iqbal e Xining (Ning) Wang. "Enhanced Diagnosis of Thyroid Diseases Through Advanced Machine Learning Methodologies". Em: *Sci* 7.2 (2025), p. 66. DOI: 10.3390/sci7020066.