Síndrome do doente eutireoidiano: análise de indicadores importantes com *machine learning*

Vinicius A. Almeida
Tecnologia da Informação
Laboratorio de Inteligencia Computacional
Pau dos Ferros RN, Brazil
vinicius.almeida37366@alunos.ufersa.edu.br

Rosana C. B. Rego
Departamento de Engenharias e Tecnologia
Laboratorio de Inteligencia Computacional
Pau dos Ferros RN, Brazil
rosana.rego@ufersa.edu.br

***Abstract*—A seleção de quais atributos utilizar é uma etapa fundamental no processo de aprendizado de máquina, pois ao selecionar os atributos adequados evitamos sobrecarregar o modelo com informações desnecessárias para fazer previsões. Neste estudo, exploramos três métodos para selecionar atributos: Eliminação Aleatória de Atributos (*Random Feature Elimination*-RFE), método de correlação e seleção baseada em agrupamento (*K-Means*). Levamos em consideração esses métodos para decidir quais atributos usar, com base no conjunto de dados fornecido pela Universidade da Califórnia.**

***Index Terms*—Tireoide, Inteligencia Artificial, Features, Classificação, Saúde.**

I. Introdução

Doenças e desordens relacionadas à tireoide são questões hormonais amplamente difundidas que afetam grande parte da população global. As doenças e distúrbios da tireoide incluem tireoidite e câncer da tireoide. A glândula tireoide é uma das glândulas endócrinas mais fáceis de distinguir, localizadas na frente da garganta e ao redor da traqueia [1].

A seleção de características desempenha um papel crucial para evitar a inclusão excessiva de atributos que não são relevantes para o processo de aprendizado de máquina. O objetivo deste processo é escolher os atributos mais significativos, aumentando assim a precisão e a clareza das hipóteses geradas pelos algoritmos. No entanto, para realizar a seleção das características, é necessário determinar a importância dos atributos com base em critérios específicos. Alguns desses critérios envolvem distância, como o *k-means*, consistência ou informações [2].

Além disso, é importante estimar o desempenho dos algoritmos de seleção de características, uma vez que não é possível determinar de antemão qual algoritmo será superior a outro. Geralmente, essas comparações são feitas ao analisar a curva de erro gerada pelo modelo a partir dos subconjuntos de características selecionadas. No entanto, outros fatores também precisam ser considerados, como a percentagem de redução das características nesses subconjuntos [2].

Outros pesquisadores também empregaram técnicas de aprendizado de máquina para prever doenças da tireoide, como demonstrado por [2]. Eles alcançaram uma precisão de aproximadamente 95.87% usando a abordagem de Rede Neural Artificial (RNN). Por outro lado, [3] utilizou a técnica de árvore de decisão e alcançou cerca de 98.43% de precisão, após remover 3 características específicas: consulta por tiroxina, consulta por hipotireoidismo e consulta por hipertireoidismo. Essa remoção foi baseada nos estudos de [4].

Neste estudo, além de investigar e analisar métodos para a seleção de atributos, exploramos três abordagens distintas: a Eliminação Aleatória de Atributos (RFE), a utilização do método de correlação e a seleção fundamentada em agrupamento (*Clustering*). Esses métodos foram aplicados com o propósito de aprimorar a compreensão da Síndrome do Doente Eutireoidiano e otimizar o processo de seleção de atributos pertinentes a essa condição. Foram escolhidos 4 métodos levando em consideração que eles são os mais utilizados, como podemos observar em outros trabalhos como em [5] e [6] e uma variação de 5 k-vizinhos mais próximos em todos os modelos.

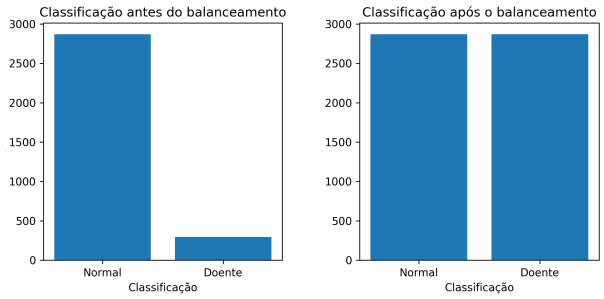
II. Conjunto de Dados

O conjunto de dados empregado foi fornecido pela Universidade da Califórnia e é composto por diversas características, tais como idade, sexo, utilização de tiroxina, histórico de consultas relacionadas à tiroxina, uso de medicação antitireoidiana, histórico de cirurgia de tireoide, consultas por hipotireoidismo, consultas por hipertireoidismo, estado de gravidez, condição de doença, presença de tumor, uso de lítio, presença de bócio, além dos níveis de TSH, T3, TT4, T4U e FTI, conforme mostrado na Tabela I.

O conjunto de dados inicial apresentava desequilíbrio entre as classes e colunas com valores ausentes. Para tratar esses desafios, realizamos procedimentos de limpeza de dados. Essas etapas incluíram a eliminação das colunas com valores faltantes, além da aplicação de técnicas como preenchimento por média e moda. Para lidar com o desequilíbrio, adotamos a técnica de balanceamento chamada SMOTE, que gera dados adicionais para a classe minoritária. A geração de novos dados pelo método SMOTE foi baseada na abordagem k-ésimo vizinho mais próximo (*K-Nearest Neighbors Algorithm* - KNN) comforme mostrado na Figura 1. Após a conclusão desse processo de preparação, os dados estão prontos para serem incorporados ao modelo.

Tabela I: *Dataset* 5 primeiras linhas.

Classificação	Idade	Sexo	Doente	Tumor	Lítio	goitre	Bócio	T3	TT4	T4U	FTI
1	45	0	0	0	0	0	1.9	1	82	0.73	112
1	64	0	1	0	0	0	0.09	1	101	0.82	123
1	56	1	0	0	0	0	0	0.8	76	0.77	99
1	78	0	0	0	0	0	2.6	0.3	87	0.95	91
1	80	1	0	0	0	0	1.4	0.8	105	0.88	120



(a) (b)

Fig. 1: (a) O conjunto de dados original, antes de passar pelo processo de balanceamento e (b) o conjunto de dados após ter sido submetido ao processo de balanceamento.

III. Técnicas de seleção de features

O processo de seleção de características é essencial para garantir que o conjunto de dados não seja treinado com características que sejam irrelevantes, ruidosas ou redundantes para o processo de classificação. Esse método desempenha um papel crucial em aumentar a precisão e reduzir o tempo necessário para a execução [7].

Dessa forma, o modelo apresentará um desempenho aprimorado e ocupará menos espaço, uma vez que não utilizará todas as características. Isso nos permitirá obter uma compreensão mais profunda dos dados, pois poderemos analisar a importância das características de maneira mais detalhada [8].

A. Seleção de Recursos de Forma Recursiva

A Eliminação Aleatória de Recursos ou RFE é um algoritmo que aprimora a seleção de características. Ele treina um modelo e elimina as características que têm a menor contribuição até que um número específico de características seja alcançado. As características selecionadas pelo RFE são aquelas que mais significativamente auxiliam na realização precisa de previsões [9].

Essa técnica assegura uma separação adequada dos dados, especialmente quando estes são de alta dimensionalidade. Os valores das características são obtidos por meio de um kernel linear, e seus pesos são empregados para avaliar quais características devem ser escolhidas. Em conformidade com [10], essa abordagem de seleção de características envolve identificar variáveis com maior poder de discriminação entre as classes, usando um método de eliminação sequencial fundamentado no princípio de maximização das margens. Isso implica treinar um classificador com um conjunto inicial de características e ir removendo-as gradualmente até atingir um limite predefinido (N), que corresponde ao número especificado de características, conforme mencionado em [11].

B. Seleção de atributos baseada na correlação

A técnica de seleção de características baseada na correlação (CFS) é conhecida por sua rapidez e capacidade de identificar características redundantes, irrelevantes e que introduzem ruído. Ela revela características independentes que não dependem de outras. Comumente, essa abordagem elimina mais da metade das características do conjunto de dados, frequentemente resultando em um aumento na precisão ao comparar com o uso de todas as características originais, conforme explicado por [12]. A seleção das características ocorre ao criar subconjuntos aninhados, começando com um subconjunto altamente correlacionado e progredindo para os menos correlacionados. Esses subconjuntos são avaliados através de validação cruzada, conforme detalhado em [13].

C. Seleção de features baseada em cluster

A técnica de seleção baseada em agrupamento envolve dividir um conjunto de características em grupos com base em critérios específicos, com o objetivo de agrupar características semelhantes em um mesmo *cluster*. Para alcançar isso, comumente se utiliza o agrupamento que maximiza a similaridade dentro dos *clusters* e a dissimilaridade entre *clusters*, como discutido por [12]. Nesse contexto, empregamos o Algoritmo de Agrupamento *K-Means*, que requer uma matriz com M pontos em N dimensões e uma matriz de K pontos iniciais para os *clusters* no centro, também em N dimensões. O número de *clusters* é determinado pela distância euclidiana. O objetivo padrão é encontrar a partição ideal de K *clusters* e a soma dos quadrados movendo de um ponto para outro dentro do *cluster*, conforme explicado por [14].

IV. Discussão e resultados

Na abordagem de seleção por correlação, conforme mostrado na Figura 2, escolhemos características com base naquelas que possuem maior correlação entre si. Nesse processo, a diagonal principal é excluída, uma vez que possui uma correlação forte consigo mesma.

Conforme mostrado nas Figuras 3 (a), (b), (c) e (d), são apresentadas seleções de 5, 8, 10 e 12 características, respectivamente. Essas representam as características escolhidas por meio do método de Eliminação Recursiva de Características (RFE), o qual seleciona características de maneira iterativa com base em um modelo algorítmico.

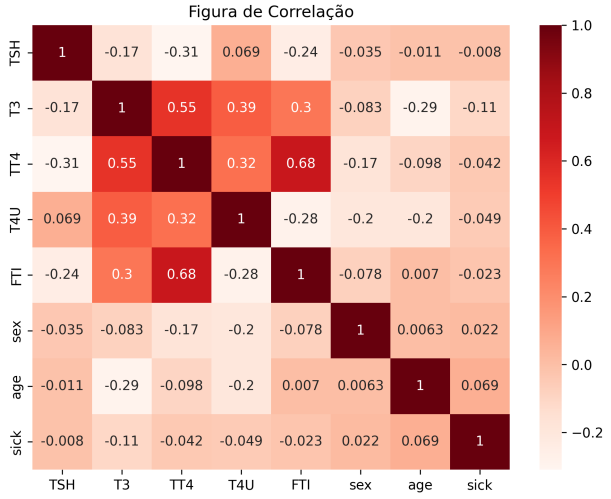


Fig. 2: Matriz de correlação.

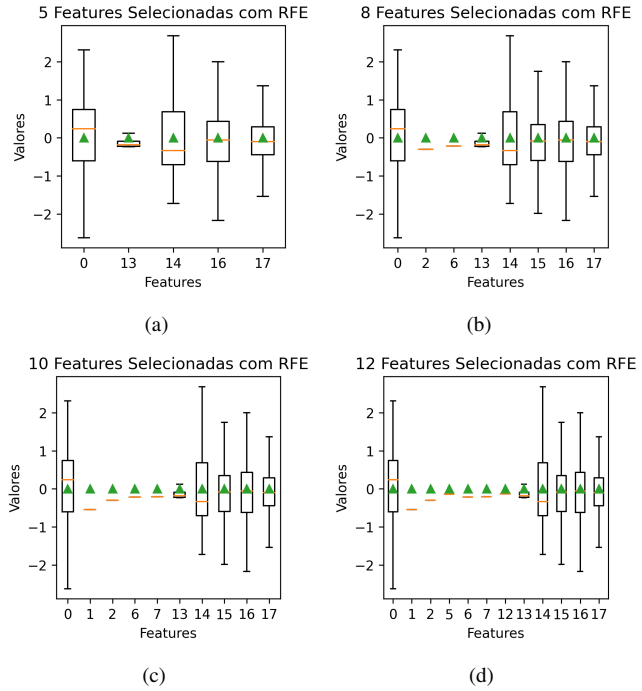


Fig. 3: Método de seleção de características recursivas: (a) 5 características, (b) 8 características, (c) 10 características e (d) 12 características selecionadas

Nas Figuras 3 (a), (b), (c) e (d), cada característica é associada a um valor específico. A correspondência entre o valor e a característica é detalhada na Tabela II correspondente.

No processo de seleção de características por agrupamento, as características são agrupadas em *clusters* contendo 3, 5, 8, 10 e 12 características relacionadas entre si, como indicado na Tabela V. A ideia é que as características mais relevantes sejam aquelas que aparecem com maior frequência nas diferentes técnicas de seleção de características utilizadas.

No método de correlação, 4 características foram selecionadas, enquanto o RFE identificou 5 características como principais. Por sua vez, o método de agrupamento selecionou 7 características. As características escolhidas para a previsão foram aquelas que apresentaram maior frequência nas três abordagens, conforme indicado na Tabela III.

Tabela II: *Features* selecionadas pelo RFE e seus respectivos números.

Feature	Número
Idade	0
Sexo	1
Usando tiroxina	2
Cirurgia de tireoide	5
Consulta hipotireoidismo	6
Consulta hipertireoidismo	7
Bócio	12
TSH	13
T3	14
TT4	15
T4U	16
FTI	17

Tabela III: *Features* selecionadas pelos métodos Correlação, RFE e Clustering.

Método	Features selecionadas
Correlação	FTI, TT4, T4U, T3
RFE	Idade, TSH, T3, T4U, FTI
Clustering	TSH, Sexo, Idade, Usando tiroxina, Consulta hipotireoide, Litio, T3
Mais aparecem	FTI, T4U, T3, TSH, Idade

Na Tabela IV, são apresentados os métodos e os modelos que demonstraram melhor desempenho para a previsão utilizando os atributos apresentados na Tabela III. Em todos os métodos de seleção de características, o modelo escolhido para conduzir as previsões foi o LightGBM. Conforme a tabela revela, as características selecionadas pelo método de correlação alcançaram uma acurácia de 0.9789, uma área sob a curva (AUC) de 0.9962, um recall de 0.9848 e um F1-score de 0.9790. Por outro lado, o método RFE alcançou uma acurácia de 0.9871, AUC de 0.9977, recall de 0.9895, precisão de 0.9849 e F1-score de 0.9872. Finalmente, o método de agrupamento resultou em uma acurácia de 0.9800, AUC de 0.9967, recall de 0.9769, precisão de 0.9830 e F1-score de 0.9799. Com base nesses resultados, conclui-se que o método RFE é o mais eficaz, já que apresentou as métricas mais elevadas. Para essa tabela ser gerada foi utilizado o método pycaret, o pycaret é um framework de python que utiliza de dezenas de algoritmos prontos para serem comparados através de diversas métricas. Ele foi setado com as features selecionadas por cada método e feito a limpeza e o balanceamento dos dados, da qual setamos como 80% dos dados para treino e 20% para teste, de forma que foram feitos teste com 14 algortimos de predição e em todas as formas de seleção de features, o melhor modelo para ser treinado foi o LightGBM, por ter as métricas mais elevadas em comparação com os outros modelos.

Tabela IV: Métricas para o modelo LightGBM em cada método de seleção de *feature*.

Método	Acurácia	AUC	Recall	Precisão	F1
Correlação	0.9789	0.9962	0.9848	0.9734	0.9790
RFE	0.9871	0.9977	0.9895	0.9849	0.9872
Cluster	0.9800	0.9967	0.9769	0.9830	0.9799

V. Conclusão

A aplicação desses diversos métodos de seleção de características é uma etapa crucial para obter um número apropriado de características, evitando excessos. Isso contribui para uma previsão mais precisa e compreensível. Como não há maneira definitiva de determinar qual método de seleção é superior, este estudo empregou várias abordagens. A conclusão que podemos tirar deste conjunto de dados é que, ao utilizar a técnica de agrupamento, as características mais importantes incluem TSH, sexo, idade, uso de tiroxina, histórico de consulta por hipotireoidismo, uso de lítio, níveis de T3, estado de gravidez, níveis de TT4, condição de doença, uso de medicação antitireoide, histórico de consulta por tiroxina, níveis de T4U, histórico de cirurgia de tireoide e histórico de consulta por hipertireoidismo.

Por outro lado, ao empregar o método RFE, as características selecionadas são idade, sexo, uso de tiroxina, histórico de cirurgia de tireoide, histórico de consulta por hipotireoidismo, histórico de consulta por hipertireoidismo, presença de bócio, níveis de TSH, T3, TT4, T4U e FTI.

Ao analisar a figura de correlação, as características destacadas são FTI, TT4, T4U e T3. Considerando as características que mais aparecem em todos esses métodos, conclui-se que as mais importantes são FTI, TT4, T4U, T3, sexo e idade.

Agradecimentos

A PICI/UFERSA pelo apoio financeiro em forma de bolsa de iniciação cientifica.

References

[1] S. S. Islam, M. S. Haque, M. S. U. Miah, T. B. Sarwar, and R. Nugraha, "Application of machine learning algorithms to predict the thyroid disease risk: an experimental comparative study," *PeerJ Computer Science*, vol. 8, p. e898, 2022.

[2] H. D. Lee, "Seleção de atributos importantes para a extração de conhecimento de bases de dados," Ph.D. dissertation, Universidade de São Paulo, 2005.

[3] E. Sonuç *et al.*, "Thyroid disease classification using machine learning algorithms," in *Journal of Physics: Conference Series*, vol. 1963, no. 1. IOP Publishing, 2021, p. 012140.

[4] I. Ioniţă and L. Ioniţă, "Prediction of thyroid disease using data mining techniques," *BRAIN. Broad Research in Artificial Intelligence and Neuroscience*, vol. 7, no. 3, pp. 115–124, 2016.

[5] B. F. Darst, K. C. Malecki, and C. D. Engelman, "Using recursive feature elimination in random forest to account for correlated variables in high dimensional data," *BMC genetics*, vol. 19, no. 1, pp. 1–6, 2018.

[6] K. K. Nicodemus and J. D. Malley, "Predictor correlation impacts machine learning algorithms: implications for genomic studies," *Bioinformatics*, vol. 25, no. 15, pp. 1884–1890, 2009.

[7] A. U. Haq, J. P. Li, M. H. Memon, S. Nazir, and R. Sun, "A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms," *Mobile information systems*, vol. 2018, pp. 1–21, 2018.

[8] K. Yan and D. Zhang, "Feature selection and analysis on correlated gas sensor data with recursive feature elimination," *Sensors and Actuators B: Chemical*, vol. 212, pp. 353–363, 2015.

[9] M. H. Memon, J. P. Li, A. U. Haq, M. H. Memon, and W. Zhou, "Breast cancer detection in the iot health environment using modified recursive feature selection," *wireless communications and mobile computing*, vol. 2019, pp. 1–19, 2019.

[10] H. Jeon and S. Oh, "Hybrid-recursive feature elimination for efficient feature selection," *Applied Sciences*, vol. 10, no. 9, p. 3211, 2020.

[11] R. Andreola and V. Haertel, "Seleção de variáveis em imagens hiperespectrais para classificadores svm."

[12] M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, The University of Waikato, 1999.

[13] I. Guyon *et al.*, "Practical feature selection: from correlation to causality," *Mining massive data sets for security: advances in data mining, search, social networks and text mining, and their applications to security*, pp. 27–43, 2008.

[14] J. A. Hartigan, M. A. Wong *et al.*, "A k-means clustering algorithm," *Applied statistics*, vol. 28, no. 1, pp. 100–108, 1979.

Tabela V: *Clusters* e as *features* selecionadas.

Clusters	2 features	3 features	5 features	8 features	10 features	12 features
3	sexo, TSH	classificação, sexo, TSH	classificação, sexo, consulta hipotireoidismo, TSH, T3	classificação, sexo, consulta hipotireoidismo, lítio, TSH, T3, TT4, T4U	classificação, idade, sexo, em medicação antitireoidiana, consulta hipotireoidismo, lítio, TSH, T3, TT4, T4U	classificação, idade, sexo, consulta sobre tiroxina, em medicação antitireoidiana, cirurgia de tireoide, consulta hipotireoidismo, lítio, TSH, T3, TT4, T4U
5	doente, lítio	doente, lítio, TSH	sexo, consulta sobre tiroxina, doente, lítio, TSH	classificação, sexo, consulta sobre tiroxina, doente, lítio, TSH, T3, TT4	classificação, sexo, consulta sobre tiroxina, consulta hipotireoidismo, doente, lítio, TSH, T3, TT4, T4U	classificação, sexo, consulta sobre tiroxina, em medicação antitireoidiana, cirurgia de tireoide, consulta hipotireoidismo, doente, lítio, TSH, T3, TT4, T4U
8	gravida, TSH	usando tiroxina, gravida, TSH	classificação, sexo, usando tiroxina, gravida, TSH	classificação, sexo, usando tiroxina, gravida, TSH, T3, TT4, T4U	classificação, sexo, usando tiroxina, usando medicação antitireoidiana, consulta hipotireoidismo, gravida, TSH, T3, TT4, T4U	classificação, sexo, usando tiroxina, em medicação antitireoidiana, consulta hipotireoidismo, gravida, doente, lítio, TSH, T3, TT4
10	idade, usando tiroxina	idade, usando tiroxina, doente	idade, usando tiroxina, gravida, doente, TT4	classificação, idade, usando tiroxina, consulta usando tiroxina, gravida, doente, TSH, TT4	classificação, idade, sexo, usando tiroxina, consulta usando tiroxina, gravida, doente, TSH, T3, TT4	classificação, idade, sexo, usando tiroxina, consulta usando tiroxina, cirurgia de tireoide, consulta hipertireoidismo, gravida, doente, TSH, T3, TT4
12	idade, consulta hipotireoidismo	idade, em medicação antitireoidiana, consulta hipotireoidismo	idade, em medicação antitireoidiana, consulta hipotireoidismo, consultar hipertireoidismo, lítio	idade, sexo, usando tiroxina, em medicação antitireoidiana, consulta hipotireoidismo, consulta hipertireoidismo, gravida, lítio	idade, sexo, usando tiroxina, em medicação antitireoidiana, cirurgia de tireoide, consulta hipotireoidismo, consulta hipertireoidismo, gravida, lítio, T3	idade, sexo, usando tiroxina, consulta sobre tiroxina, em medicação antitireoidiana, cirurgia de tireoide, consulta hipotireoidismo, consulta hipertireoidismo, gravida, lítio, T3, T4U