Ablation dosage recommendation for thyroid cancer treatment following thyroidectomy using machine learning

Lai Phu Minh[1], Pham Thanh Vinh[2], Pham Duc Thuc[2], Nguyen Thanh Trung[3], Tran Quoc Long[2], Nguyen Thi Phuong[3], Chu Minh Duc[3], Tran Van Dien[3], Pham Ha Hai[3], Nguyen Thai Ha[1], Nguyen Duc Thuan[1], Nguyen Chi Thanh[4*]

[1]Country Department of Electronics, School of Electrical and Electronic Engineering, Hanoi University of Science and Technology, 1 Dai Co Viet, Hai Ba Trung, Hanoi, Vietnam;
[2]Institute for Artificial Intelligence, VNU University of Engineering and Technology, 144 Xuan Thuy, Cau Giay, Hanoi, Vietnam;
[3]108 Military Central Hospital, 1B Tran Hung Dao, Hai Ba Trung, Hanoi, Vietnam;
[4]Institute of Information Technology, Academy of Military Science and Technology, 17 Hoang Sam, Cau Giay, Hanoi, Vietnam.
[*]Corresponding author: thanhnc@ioit.ai.vn
Received 1 May 2024; Revised 16 Apr. 2024; Accepted 12 Jun. 2024; Published 25 Jun. 2024.
DOI: https://doi.org/10.54939/1859-1043.j.mst.96.2024.137-144

ABSTRACT

This article presents an innovative approach to ascertain the most effective ablation dosages for thyroid cancer treatment following thyroidectomy. The methodology utilizes Decision Trees and places significant emphasis on the interpretability of medical decision-making. By incorporating clinical data and the Radioactive Scan Index (RSI) into Decision Tree algorithms, our methodology offers transparent treatment planning insights. By means of a case study, we illustrate the function of Decision Trees in clarifying pivotal elements that impact dosage recommendations for ablation, thereby enabling medical practitioners to make well-informed decisions. This study emphasizes the importance of decision explainability in the optimization of treatment strategies for thyroid cancer, ultimately leading to enhanced patient care and treatment outcomes.

Keywords: Thyroid treatment; Machine learning; Decision tree.

1. INTRODUCTION

Thyroid cancer, particularly the prevalent papillary subtype, poses significant challenges due to its ability to exhibit rapid progression and resistance to conventional treatments. Early detection remains a critical factor in enhancing patient prognosis, yet the absence of symptoms in the initial stages complicates timely diagnosis. Current therapeutic strategies, including thyroid hormone therapy, radioactive iodine treatment, and surgery, often require precise customization to optimize outcomes and mitigate long-term impacts on patients' lives. This research introduces a novel machine learning-based approach to refine ablation dosage recommendations post-thyroidectomy, which is crucial for effective thyroid cancer management. Utilizing Decision Tree algorithms integrated with clinical data and the Radioactive Scan Index (RSI), this study aims to enhance the precision and transparency of treatment planning. The decision-making process in medical settings demands high interpretability to ensure that practitioners can make informed and confident treatment decisions. The paper is structured as follows: Section 2 details the materials and methods, including the specific machine-learning tools and data sets used in our analysis. Section 3 presents the results of our model's application to a cohort of thyroid cancer patients, illustrating its effectiveness in dosage prediction and its potential impact on patient outcomes. Section 4 discusses these findings in the broader context of thyroid cancer treatment, emphasizing the practical implications and potential for future research and application. Finally, the conclusion summarizes the contributions of our study and outlines avenues for further investigation. By systematically addressing these aspects, this introduction sets the stage for a comprehensive exploration of how advanced data-driven techniques can revolutionize the management of thyroid cancer post-surgery.

2. MATERIALS AND METHOD

2.1. Software requirements

With a primary focus on machine learning and data visualization, the Python 3.9 programming language is mandated for the execution of this system's software requirements. The software establishes Decision Trees (DTs) to facilitate the development of resilient models by utilizing the sklearn program. Graphviz [12] is utilized to improve the visual representation of these structures. In addition, the seaborn package aids in the visual representation of data, whereas the agile pandas package is utilized to facilitate efficient data manipulation and processing. Compatibility, functionality, and an intuitive user experience are all ensured by the exhaustive environment formed by these software components.

2.2. Data

The data utilized in this research were obtained from a sample of 3213 patients diagnosed with thyroid cancer who sought imaging services at the 108 Military Central Hospital in Hanoi, Vietnam, between March 2018 and March 2021. The patient population consisted of 2684 females and 529 males. The age distribution of the individuals was one-over-five males to females, spanning from 18 to 72 years. All patients received medical care in accordance with the 108 Military Central Hospital's standardized protocol. Certain patients with distal metastases received either the initial or subsequent treatment (i.e., those who exhibited residual thyroid tissues following the initial treatment). Conversely, others undertook a reevaluation that confirmed the absence of residual thyroid tissues. Furthermore, the patients exhibited a diverse range of thyroid cancer subtypes, including variant forms, cystic, and nodular. Nevertheless, the scope of this research is limited to patients who are undergoing their initial treatment for nodular thyroid cancer and still have residual thyroid tissue. As a result, the dataset that remained consisted of 1477 individuals, all of whom were 18 years of age or older and had granted informed assent to partake in the research with the authorization of the Ethics Committee of 108 Military Central Hospital (number 117456).

The primary objectives of the initial examination are to ascertain the stage of the disease, evaluate risk factors, and devise a strategy for ongoing monitoring. Patient inquiries encompass pertinent personal information, surgical history, and specifics regarding cancer diagnosis. Diagnostic examinations consist of imaging studies, such as a neck ultrasound and biochemical analyses. Patients are given explicit directives regarding I-131 treatment, which may include the cessation of thyroid hormone use and the adoption of a low-iodine diet.

Initial patient inquiry and examination, pre-treatment diagnostic tests, I-131 treatment scheduling, patient preparation, and post-treatment imaging are all components of the treatment procedure. By ensuring a comprehensive assessment of each patient's condition, the integrated approach enables the development of individualized treatment programs. This article underscores the significance of patient-physician communication in treatment decision-making, specifically focusing on the alternatives to 131I therapy and follow-up examinations. The all-encompassing structure of this methodology seeks to enhance the management of thyroid cancer following thyroidectomy, establishing a structure for efficient and personalized patient attention. A statistical analysis of one quantitative and one categorical feature, out of a total of 42 features in the dataset used in this study, is presented in tables 1 and 2, respectively. As a result of the varied data types, which included categorical and quantitative fields, we conducted the subsequent data preprocessing procedures: The value range for quantitative variables was normalized to 0 to 1 through the utilization of the Min-max Scaler method [1]. With respect to categorical fields, one-hot encoding was implemented [2].

Table 1. *Statistical analysis of selecte prominent quantitative features within the dataset along with their corresponding value ranges.*

Variable	Mean ± std	Data type	Range
Tg (mmol/ml)	13.86 ± 55.45	Quant	[0, 80]
ATg (mlU/ml)	32.49 ± 89	Quant	[0, 120]
TSH (mIU/ml)	70 ± 31.8	Quant	[0, 120]

Table 2. *A statistical compilation of notable categorical features commonly employed in diagnosing ablation doses in practical medical cases, accompanied by a set of discrete possible values.*

Variable	Data type	Rage
T	Categorical	[T1, T1a, T3, T3a, ...]
M	Categorical	[M0, M1]
N	Categorical	[N0, N1, N1a, N1b]
Residual tissue	Binary	[0, 1]
Recurent risk	Categorical	TxMyNz
Pregnant	Binary	[0, 1]
Cancer stage	Categorical	[1, 2, 3, 4a, 4b]
Metastatsis	Binary	[0, 1]
Dose (MCI)	Categorical	[50, 75, 100]

2.3. Proposed model

Decision trees are preferred in scientific settings for regression and classification tasks owing to their visual simplicity and interpretability. In order to operate, these models partition data into homogeneous subsets according to classification and regression criteria, respectively, such as Gini impurity or entropy. This procedure terminates when particular halting conditions, such as minimal node size or maximum depth, are satisfied; it employs recursive binary splitting. Pruning techniques have been implemented to prevent overfitting and improve the dependability of models.

In addition to effectively handling non-linear relationships and complex feature interactions, decision trees exhibit resilience in the face of outliers and absent values. They play a crucial role in the identification of significant predictive variables, thereby adding value to the process of data exploration. Nevertheless, in order to mitigate the risk of overfitting, particularly when dealing with intricate datasets, decision trees are frequently combined with ensemble techniques such as Random Forests; this compromises analytical profundity for the sake of simplicity. Because of their versatility, they are utilized in a wide range of scientific contexts.

We utilized the Decision Tree algorithm, more precisely C4.5, to accomplish the classification assignment in our research. In order to maximize the efficiency of the model, we employ a grid search methodology to refine hyperparameters within the specified range, as outlined in table 3. By utilizing this grid search methodology, we are able to systematically evaluate an extensive range of hyperparameter combinations and select the one that yields the highest level of performance tailored to our particular dataset. Explanations in greater detail for each parameter are available at [3]. Through an investigation of these hyperparameter ranges, our objective is to ascertain the optimal configuration for our Decision Tree C4.5 model, thereby augmenting its capacity for prediction and generalization in relation to the provided classification task.

Table 3. *Hyperparameter Grid Search Space for Decision Tree C4.5.*

Hyperparameter	Values
ccp alpha	[0.00, 0.002]
min samples leaf	[1, 2, 4, 5, 10]
min samples split	[2, 5, 10, 50]
max depth	[5, 10, 20]
min impurity decrease	[0, 0.001, 0.002, ..., 0.05]

2.4. Evaluation of the models

The assessment of a model is an essential component in the progression of predictive models, as it guarantees the models' dependability and practicality. The metrics that are frequently employed to assess classification models are delineated in this section: Precision, Recall, F1 Score, Confusion Matrix, and ROC Curve. These metrics offer valuable insights regarding the accuracy, error rate, and capacity to strike a balance between the sensitivity and specificity of a model. Particularly in machine learning and data science, assessing the efficacy of a model is equally as important as the model itself in predictive modeling. The selection of evaluation metrics is contingent upon the particular specifications and characteristics of the dataset. Various metrics are employed to evaluate the performance of classification problems, with each metric offering distinct perspectives on the merits and demerits of the model.

Precision: Precision is defined as the ratio of true positives to the sum of true and false positives. It measures the model's accuracy in classifying a sample as positive:

$$Precision = \frac{True\ Positives}{\text{True Positives} + \text{False Positives}}$$

Recall: Recall, or sensitivity, measures the proportion of actual positives that are correctly identified by the model:

$$Precision = \frac{True\ Positives}{\text{True Positives} + \text{False Negatives}}$$

F1 Score: The F1 Score is the harmonic mean of precision and recall, offering a balance between the two by considering both false positives and false negatives:

$$F1\ Score = \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Confusion Matrix: A Confusion Matrix is a table layout that allows visualization of the performance of an algorithm. It presents the counts of true positives, true negatives, false positives, and false negatives.

ROC Curve and AUC [9]: The Receiver Operating Characteristic (ROC) curve is a plot that illustrates the diagnostic ability of a binary classifier. The Area Under the Curve (AUC) provides an aggregate measure of the model's performance across all classification thresholds.

2.5. Structure of the experiment

A subset was generated from the 1600-item dataset, designating 200 samples as the test set and the remainder as the training set. Clinical data pertinent to the tumor, including blood test results, recurrence risk assessment, and tumor characteristics, were utilized in each instance to determine the optimal dosage for ablation treatment.

The C4.5 Decision Tree model was subsequently trained alongside a few other traditional classification models listed in table 4. Our objective was to determine how well they could predict the correct ablation dosage following automatic treatment, as well as the performance of the various model classes. The input for the model consisted of patients who had residual thyroid tissue, with the clinical information and RSI mentioned earlier. The output of the model represented the suggested dosages of ablation, which were 50MCI, 75MCI, and 100MCI. By employing cross-validation [4], the optimal model was determined from the 1400 training samples by identifying the variant with the highest F1 score.

The aforementioned model was subsequently employed to determine the appropriate dosage of ablation for the test set. Following that, the evaluation of the model's performance on the test set was conducted utilizing the metrics specified in section 2. In addition, we performed an analysis to determine the relative significance of every clinical information feature in an effort to identify variables that have a substantial impact on the decision-making process.

3. RESULTS AND DISCUSSION

3.1. Model performance

As shown in table 4, the Decision Tree C4.5 model performs significantly better than the others. The F1 scores are 0.82, 0.74, and 0.84, respectively, for the 50MCI, 75MCI, and 100MCI dosage levels. The superior recall exhibited by the C4.5 DT model in comparison to the other models at the 75MCI threshold is especially noteworthy. The suboptimal predictive performance of all models with respect to the 75MCI dosage level can be attributed to the relatively small quantity of data accessible for samples in this class in comparison to other classes. The aforementioned results demonstrate the model's reliable ability to identify individuals who require the specified dosage levels.

The confusion matrix generated by the C4.5 model using testing datasets is illustrated in figure 1. The computed results, which are derived from the confusion matrix, are presented in table 4. The C4.5 model exhibited a cumulative accuracy of 85% when evaluated on the test set.

In order to assess the classification capability of the model, we utilized the ROC Curve. The area under the ROC Curve is denoted by the Area Under Curve (AUC), with a greater AUC indicating that the model possesses a superior and more dependable capability for classification. ROC curves are illustrated in figure 2 to facilitate additional evaluation of model performance. The accuracy of the model in differentiating between dosage levels is illustrated by the micro and macro average AUC values. The significance of each dosage level in evaluating the discrimination capability of the model is demonstrated by the respective AUC values of 0.89 for 50MCI, 0.92 for 75MCI, and 0.89 for 100MCI in the ROC curve.

Table 4. *Comprehensive results based on evaluated metrics in both training and testing sets.*

Model	Categories	Precision	Recall	F1-score
DT-C4.5 [3]	50MCI	**0.80**	0.85	**0.82**
	75MCI	0.67	**0.83**	**0.74**
	100MCI	0.90	**0.80**	**0.84**
Nearest Neighbors [11]	50MCI	0.64	0.87	0.74
	75MCI	0.58	0.29	0.39
	100MCI	0.89	0.74	0.81
Linear SVM [10]	50MCI	0.70	0.87	0.78
	75MCI	0.65	0.62	0.64
	100MCI	**0.93**	0.76	0.83
MLP [5]	50MCI	0.70	0.87	0.78
	75MCI	**0.80**	0.50	0.62
	100MCI	0.88	0.79	0.83
AdaBoost [6]	50MCI	0.66	**0.88**	0.75
	75MCI	0.67	0.33	0.44
	100MCI	0.90	0.77	0.83

3.2. Feature importance

The purpose of this experiment is to determine the degree of interdependence between the features and the ablation dosage predictions made by the model; alternatively, it seeks to determine the importance of each feature with respect to diagnostic outcomes. The permutation feature importance method [6] is implemented, with the f1 macro average score serving as the metric. Determining the significance of each feature involves quantifying the decrease in the metric.

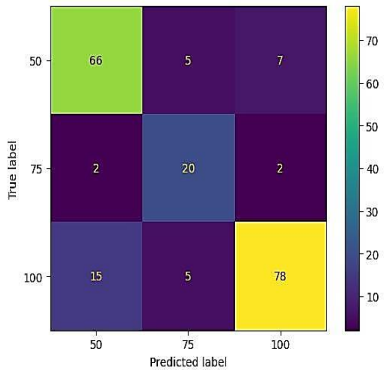

Figure 1. *Confusion matrix of C4.5 model computed on the test set.*

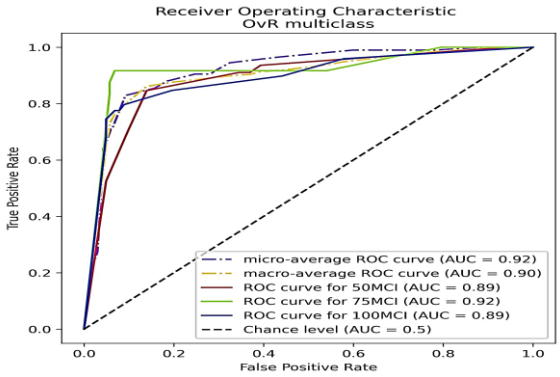

Figure 2. *Evaluated ROC curve over test set.*

When this specific feature is perturbed while the remaining features remain unchanged, this process is repeated in numerous permutations.

The box plot in figure 3 illustrates the importance of the characteristics utilized by our proposed model for the purpose of diagnosing 131I ablation dosages. The results indicate that while the model incorporates 42 features for the purpose of diagnosis, only a limited subset of these features proves to be operational. Figure 4, which visually represents our proposed model in the form of a tree, substantiates this assertion. These outcomes indicate that the model exclusively employs the input features in practice. This result is consistent with the pragmatic observations made by medical professionals while managing thyroid carcinoma at 108 Military Central Hospital.

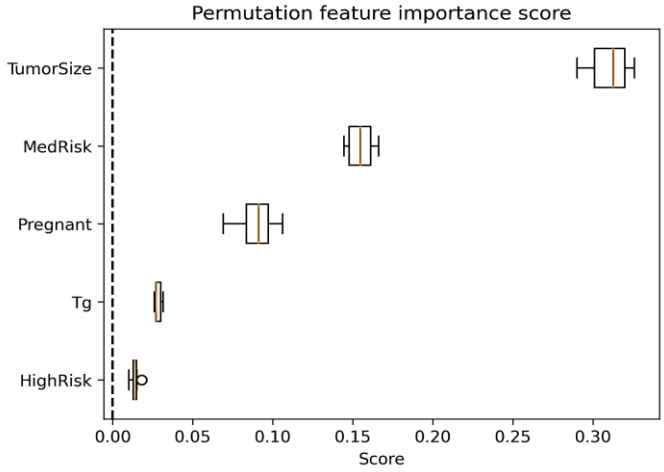

Figure 3. *The top 5 influential features that significantly impact our proposed model's decision-making regarding dosage determination.*

3.3. Discussion

The present research examined the viability of employing a Decision Tree C4.5 model to aid in the determination of 131I dosage for patients who have thyroidectomy but still have residual

thyroid tissue. With an overall accuracy of 85% on the test set, the model effectively classified patients into the following three dosage categories: 50MCI, 75MCI, and 100MCI. Additionally, the decision tree model's ability to be interpreted improves its practicality as a supplementary instrument for medical practitioners throughout the duration of treatment. This characteristic allows for concise elucidations of the model's deliberative processes, thereby facilitating its potential incorporation as a supplementary tool for healthcare practitioners. Potential avenues for future advancement may encompass:

- **Expanding the dataset:** Enhancing the diversity and representation of the dataset by collecting more samples from various sources or utilizing data from multiple databases to improve the model's generalizability.
- **Enhancing model mobility:** Developing a version of the model deployable on mobile or online platforms, enabling physicians to conveniently access and utilize the model during patient care. Combining various machine learning approaches.

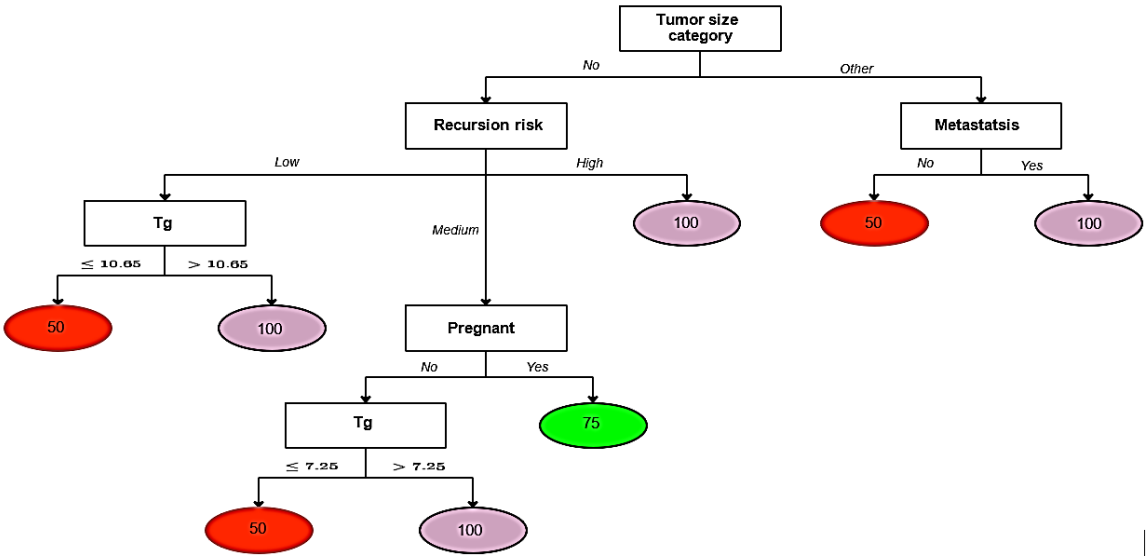

Figure 4. *The decision tree represents the model.*

- Integrating multiple machine learning methods, such as neural networks, deep learning, or other machine learning models, to leverage the strengths of each method enhances prediction performance and model accuracy.
- Clinical validation: Conducting trials and clinical validation on a larger dataset and performing real-world experiments to verify the applicability and reliability of the model in practical settings.

4. CONCLUSIONS

To summarize, our research examined the utilization of machine learning techniques in recommending 131I dosages for the purpose of treating thyroid cancer subsequent to thyroidectomy. The model that was developed demonstrated encouraging discriminatory abilities when it came to determining suitable dosage levels using clinical data. The potential utility of the Decision Tree model as a supplementary instrument for clinicians is attributed to its interpretable character. Additional research in the areas of dataset expansion, model mobility, and clinical validation may significantly augment the practical utility of this approach and make a positive contribution to the quality of patient care.

Acknowledgement: *This research has been done under the research project TXTCN.21.23 of Vietnam National University, Hanoi.*

REFERENCES

[1]. Patro, S.; Sahu, K. K. *"Normalization: A preprocessing stage"*. arXiv preprint arXiv:1503.06462, (2015),

[2]. Hancock, J. T.; Khoshgoftaar, T. M. *"Survey on categorical data for neural networks"*. Journal of Big Data, 7, 1–41, (2020).

[3]. Scikit-learn: Decision Tree Classifier. https://scikitlearn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html#sklearn.tree.DecisionTreeClassifier.

[4]. Campbell, S. L.; Gear, C. W. *"The index of general nonlinear DAES"*. Numer. Math., 72, 173–196, (1995).

[5]. Murtagh, F. *"Multilayer perceptrons for classification and regression"*. Neurocom- puting, 2, 183–197, (1991).

[6]. Freund, Y.; Schapire, R.; Abe, N. *"A short introduction to boosting"*. Journal- Japanese Society For Artificial Intelligence, 14, 1612, (1999).

[7]. Browne, M. W. *"Cross-validation methods"*. Journal of mathematical psychology, 44, 108–132, (2000).

[8]. Breiman, L. Random forests. *"Machine learning"*, 45, 5–32, (2001).

[9]. Bradley, A. *"The use of the area under the ROC curve in the evaluation of machine learning algorithms"*. Pattern recognition, 30(7), 1145–1159, (1997).

[10]. Auria, L., & Moro, R. *"Support vector machines (SVM) as a technique for solvency analysis",* (2008).

[11]. Mucherino, A., Papajorgji, P., Pardalos, P., Mucherino, A., Papajorgji, P., & Pardalos, P. *"K-nearest neighbor classification"*. Data mining in agriculture, 83–106, (2009).

[12]. Ellson, J., Gansner, E., Koutsofios, E., North, S., & Woodhull, G. *"Graphviz and dynagraph—static and dynamic graph drawing tools"*. Graph drawing software, 127–148, (2004).

[13]. Bryan R Haugen et al *"2015 American Thyroid Association Management Guidelines for Adult Patients with Thyroid Nodules and Differentiated Thyroid Cancer: The American Thyroid Association Guidelines Task Force on Thyroid Nodules and Differentiated Thyroid Cancer"* PMID: 26462967 PMCID: PMC4739132 DOI: 10.1089/thy.2015.0020.

TÓM TẮT

Sử dụng học máy để khuyến cáo về liều điều trị ung thư tuyến giáp sau phẫu thuật cắt toàn bộ tuyến giáp

Một cách tiếp cận được trình bầy rõ ràng đầy sáng tạo trong bài viết này để xác định đề xuất liều lượng dược chất phóng xạ hiệu quả nhất trong điều trị ung thư tuyến giáp sau phẫu thuật cắt bỏ tuyến giáp. Phương pháp này sử dụng Cây quyết định và nhấn mạnh đáng kể vào khả năng diễn giải của việc ra quyết định tại các cơ sở y tế. Bằng cách kết hợp dữ liệu lâm sàng và chỉ số (RSI) vào thuật toán cây quyết định, phương pháp của chúng tôi cung cấp thông tin chi tiết về kế hoạch điều trị một cách minh bạch. Bằng một nghiên cứu điển hình, chúng tôi minh họa chức năng của Cây quyết định trong việc làm rõ các yếu tố then chốt ảnh hưởng đến đề xuất về liều lượng điều trị sau phẫu thuật cắt bỏ tuyến giáp, từ đó cho phép các bác sĩ đưa ra quyết định một cách tự tin hơn. Nghiên cứu này nhấn mạnh tầm quan trọng của khả năng giải thích quyết định trong việc tối ưu hóa các phác đồ điều trị ung thư tuyến giáp, cuối cùng dẫn đến kết quả chẩn đoán và điều trị bệnh nhân ung thư tuyến giáp được nâng cao.

Từ khoá: Điều trị tuyến giáp; Học máy; Cây quyết định.