Contents lists available at *Dergipark*

Journal of Scientific Reports-A

journal homepage: *https://dergipark.org.tr/tr/pub/jsr-a*

Analysis of patient demographics and test results using data mining methods and thyroid cancer examination

Şükrü Kitiş[a,*]

[a] *Simav Technology Faculty, Kütahya Dumlupınar University, Kütahya, 43500, Turkey*

Abstract

The aim of this study is to detect thyroid cancer and cancer recurrence using 383 datasets containing 16 parameters. These variables are: age, gender, smoking, smoking history, whether got radiotherapy or not, thyroid function status, physical examination, adenopathy, pathology, focus, risk type, T, N, M stages depending on risk type, cancer level and recurrence status. In this study, Decision Stump, Hoeffding Tree, J48, LMT, Random Forest1, Random Forest2, REP Tree trees datasets and Naive Bayes, Logistic Function, Multilayer Perception Function, Simple Logistic Function1, Simple Logistic Function2, IBK K 3 functions were run with WEKA program. According to the results, it is concluded that Random Forest trees are better than other classifiers and studies in the literature.

Keywords: Thyroid, WEKA, Data Mining

1. Introduction

With the developing technology, smart systems and computer systems becoming widespread today, many raw data have been stored. The meaningful data in this big and meaningless data, which does not make sense on its own, needs to be made meaningful in a way that will meet the needs of future generations and can be used for the benefit of humanity together with the existing data. Data mining has gained a place today as an interdisciplinary working model that has proven to be reliable with its structure that supports those who analyze in this field, provides convenience and ideas to decision-makers, and provides an opportunity to take precautions before problems occur [1].

The health sector is also a sector where raw data is available. For this reason, data mining professionals have conducted many studies in the health sector. For example, Health Information Systems provide ease of decision-making for physicians in the diagnosis and treatment of disease. Thanks to these systems, data flow is provided in areas such as quality management, diagnosis and treatment of disease, medical documentation and information management [1].

*Corresponding author. Tel.:+9-0545-334-3233; fax:+9-0274-443-0381.

E-mail address: sukru.kitis@dpu.edu.tr

In the medical field, data mining is used in many areas such as early diagnosis of diseases, treatment planning, cancer diagnosis and management of health services. Databases created from data such as blood tests, operations, external findings and physicians' opinions can be considered as a method to improve the quality of future diagnosis and treatment services by being examined by data mining algorithms. Accurate disease diagnosis reduces the time allocated for early diagnosis and treatment and the costs incurred in this process [2].

Data mining is also used in many methods used in cancer diagnosis and prognosis. In one study, a data mining model was developed for decision making for the treatment of breast cancer patients. This model helps to determine the most appropriate treatment plan using disease stage, cancer type and patient's lifestyle factors. Apart from this, there are also studies such as the detection of drug side effects, treatment recommendations, drug discovery in some diseases such as the diagnosis of diabetes, the diagnosis of Down syndrome in a baby in the womb, early diagnosis of some types of cancer, determination of the disease diagnosis of various autoimmune diseases because they do not show symptoms in the body [2]. Many studies have also been conducted on thyroid disease [3]. Hyperthyroidism is a disease caused by excessive thyroid secretion from the thyroid gland [4]. Thyroid hormones are responsible for regulating the body's energy. When thyroid hormone levels are high, the body quickly uses up energy and vital functions are accelerated. In most cases, hyperthyroidism is caused by a problem with the thyroid gland itself, and the thyroid gland simply produces an excess of thyroid hormone without any other symptoms. Rarely, hyperthyroidism can occur as a result of overproduction of thyroid stimulating hormone (TSH) from the pituitary gland [5]. As a rule, the diagnosis of hyperthyroidism is based on changes in the levels of serum thyroid hormone parameters. The thyroid hormone parameters considered in the diagnosis of hyperthyroidism are TSH, ST3, ST4, TT3 and TT4 [6]. As these enzymes are used in the diagnosis and identification of each individual person, they create huge piles of data. Furthermore, as this type of research in biology and medicine increases, it leads to a proliferation of different data types related to real-world phenomena [7]. Some of the main data types currently in use are enzymes [7], biological state measurements [8], obesity [9], DNA [10], Parkinson’s disease [11], and protein synthesis [12].

2. Method

2.1. Data set

The dataset used in this study [13] contains 16 clinicopathologic features aimed at predicting recurrence of well-differentiated thyroid cancer. The dataset has been collected over 15 years, and each patient has been followed for at least 10 years. This dataset contains 383 different data, each with 16 variables. The data was obtained from thyroid disease dataset provided by UCI Machine Learning Repository. Table 1 contains information about this dataset.

Table 1. Thyroid disease datasets provided by UCI machine learning repository

Parameter	Description	Value type
Age	Age	Year (between 15-82)
Gender	Gender	F, M (Female, Male)
Smoke	Smoking	True, false (Present, absent)
Hx smoke	Smoking history	True, false (Present, absent)
Hx radiothreapy	Radiotherapy history	True, false (Present, absent)
Thyroid	Thyroid function status	Euthyroid, clinical hyperthyroidism, others
Physical examination	Physical examination status	Multinodular goiter, single nodular goiter, others
Adenopathy	Presence of adenopathy	No, right, others
Pathology	Pathology result	Papillary, micropapillary, others
Focalization	Focus type	Multi focal, uni focal
Risk	Risk type	High, intermediate, low
T	T	T1a, T1b, T2,T3a, T3b, T4a, T4b,
N	N	N1a, N1b, N0
M	M	M0, M1

Stage	Cancer level	I, II, III, IVA, IVB
Response	Response to treatment	Indeterminate, excellent, structural incomplete, biochemical incomplete
Recurred	Cancer recurrence status	Yes, no

The list and descriptions of the attributes of the data in the Kaggle dataset in Table 1 are as follows:

Age: Age of the patient at the time of diagnosis or treatment.

Smoke: Whether the patient smokes or not.

Hx smoke: Patient's smoking history.

Hx radiotherapy: History of radiotherapy treatment.

Thyroid: The status of thyroid function.

Physical examination: Findings from physical examination of the patient.

Adenopathy: The presence or absence of enlarged lymph nodes (adenopathy).

Pathology: Specific types of thyroid cancer are determined by pathologic examination of biopsy specimens.

Focalization: Whether the cancer is unifocal (limited to a single location) or multifocal (present in more than one location).

Risk;

T: Classification of the tumor based on its size and the extent to which it has spread to nearby structures.

N: Nodal classification indicating involvement.

M: Metastasis classification indicating the presence or absence.

Staging: The overall stage of the cancer, typically determined by combining the T, N and M classifications.

Response to treatment: Indicates whether the cancer responded favorably, unfavorably or remained stable after treatment.

Recurrence status: Indicates whether the cancer is likely to recur after initial treatment [13].

Figure 1 shows the distribution of the data in the dataset according to patients.

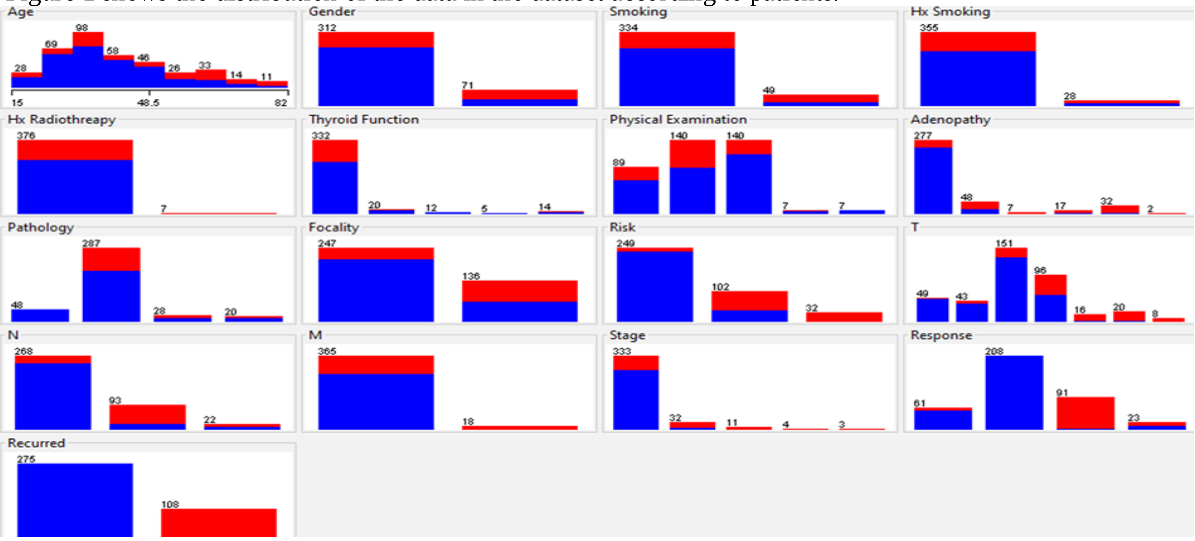


Fig. 1. Distribution of the data in the dataset according to patients (Age, gender, smoking, thyroid function etc).

Data were collected from 312 women and 71 men. The age range was between 15 to 82 years. 334 were non-smokers and 49 were smokers. 355 people had no smoking history, while 25 had a history. Other variables and their distributions are shown in Figure 1.

2.2. Algorithms

Some of the methods used in this study are as follows:

- Random Forest (RF)

RF is a classification algorithm consisting of multiple decision trees. These individual trees produce results that are then combined to provide a final prediction. Random forest is particularly useful for complex data where the relationships between variables can be very noisy. To solve such a problem, the algorithm selects a random subset of the data and builds a decision tree based on this subset. This process is repeated several times to build a set of trees. The final prediction is made by majority voting or averaging the predictions of all trees. The advantages of this algorithm are its accuracy and its ability to avoid overfitting thanks to its integrated mechanisms. It is also a versatile algorithm that can adapt to regression problems as well as classification problems. However, it can be expensive in terms of calculation and its accuracy may decrease when applied to small data sets [14].

- Naive Bayes Classification Algorithm

This algorithm is an easy-to-understand and fast method based on Bayesian theory. It is an algorithm based on the assumption that all variables are independent of each other and all have the same importance [15,16,17]. It is both a predictive and descriptive classification technique. It is one of the most effective inductive learning algorithms for machine learning and data mining. Although the independence assumption is very rare in reality, i.e. this assumption is often unrealistic, the success of Naive Bayes in classification and its superiority over some other classification algorithms has been demonstrated in various studies. The independence assumption allows each variable to be learned individually. Thus, it allows the classification process to be fast even in data with many variables. Naive Bayes classification algorithm is performed by selecting the class with the highest probability similarity to sample to be classified according to Bayes rule. A priori probabilities are used to calculate this selection. In general, 80% of the data is divided as training set and 20% as test set [16,17].

- J48 Classification Algorithm

J48 is a decision tree algorithm developed by J. R. Quinlan based on the C4.5 algorithm [18]. This algorithm tries to find appropriate behaviors for attributes with several sample approaches. Thus, it generates rules to reach the variable to be accessed. When creating decision trees, subtrees are created, and these subtrees can be further subdivided into more branches according to the dataset structure. J48 also performs pruning to remove unwanted and meaningless branches from the decision tree. The purpose of the pruning technique in J48 is to reduce misclassification errors. The decision tree consists of a decision node and a leaf node. The decision node determines the test of the features, and the leaf node determines the class values. Thus, the J48 classifier algorithm creates easy-to-understand models and provides in improvement classification performance [19,20].

- Logistic Regression (LR)

LR is a nonlinear regression model specifically designed for a binary dependent variable. It is a nonlinear model that can be linearized with appropriate transformations. In the literature, it is also called logit regression. If the dependent variable in the model is expressed with two categories, the model is called “Binary Logistic Regression Model” and if it is expressed with more than two categories, it is called “Multiple Logistic Regression Model” [21]. In the binary logistic regression model, it takes the value 1 if the event occurs and 0 if it does not occur. The independent variables in the model can be continuous and/or categorical variables, and binary or triple interactions of independent variables can be included in the model as covariates [22]. Logistic regression is one of the multivariate analysis methods used for grouping observations. In addition to its ease of use due to its lack of assumption constraints, mathematical flexibility of the model obtained from the analysis increases the interest in the method [23].

- Multilayer Perceptron Function (MLP)

MLP is a type of multilayer artificial neural. The network is also referred to as a feed-forward artificial neural system. A multilayer perceptron consists of at least three node layer components: input layer, hidden layer and output layer. It is the most widely used and best learning artificial neural network model. It is a very powerful function for classifying prediction problems. The aim of this model is to minimize the difference. The result is obtained with the target result (output) of the network [24].

- Decision Stumps (DS)

DS stands for decision logs. They are basically single-label decision trees. They are often used with big data. A log is opposed to a tree with multiple layers. It basically stops after the first split. For smaller datasets, they are also not very helpful for building simple yes/no decision models. This algorithm not only improves model performance through ensemble but also preserves the interpretability of the decision tree. It is also preferred to avoid the problem of reduced interpretability of information after integration due to the excessive depth of decision trees [25,26].

- Hoeffding Tree (HT)

T known as streamed decision tree induction. Its name comes from Hoeffding boundary used in tree induction. HT known decision tree algorithm, was initially proposed to address classification problems in large-scale data stream mining. HT model, each instance in training database is scanned once. HT possesses an outstanding computational efficiency with relatively lower RAM [25,26].

- Logistic model tree (LMT)

LMT stands for logistic model tree. It is a classification model with correlated supervision. It combines decision tree learning and logistic prediction as training algorithms. Logistic model trees use a decision tree with linear regression models at the leaves to provide a linear result on a partition basis [25,26].

- REP Tree

REP Tree is a fast decision tree learner. It prunes and builds a decision/regression tree by taking entropy as a measure of impurity and using reduced error pruning. It only ranks and qualifies numeric values once [25,26].

When it is studied in the WEKA program with these methods, parameters evaluated in terms of accuracy of the results are given below [27,28]:

- Precision

It expresses how many positive predictions are correct. Precision is especially important in situations where the cost of false alarms is high.

$$Precision = \frac{True\ Pozitives\ (TP)}{True\ Pozitives\ (TP) + False\ Pozitives\ (FP)} \tag{1}$$

- Accuracy

It is the proportion of correct predictions over the total set of examples. It is a simple and understandable measure.

$$Accuracy = \frac{TP+TN(\ True\ Negatives)}{TP+TN+FP+FN\ (False\ Negatives)} \tag{2}$$

- Recall

It is the proportion of correctly predicted observations among true positive observations. It measures the ability of the model to correctly detect all positive samples. Sensitivity is critical when false negatives are costly.

$$Recall = \frac{TP}{TP\ +FN} \tag{3}$$

- F1-score

It is the harmonic mean of precision and sensitivity. It takes the best value as 1 and the worst value as 0. The F1-score is very useful when it is necessary to find equilibrium in unbalanced datasets.

$$F1\ score\ = \frac{2\ \times Sensitivity \times Precision}{Sensitivity + Precision} \tag{4}$$

- Roc Area or Area Under the Curve (AUC)

The Receiver Operating Characteristic (ROC) curve shows the performance of the model at different thresholds and AUC represents the area under this curve.

AUC = 1.0: A perfect model (completely separates positive and negative classes).

AUC = 0.5: A random model (cannot distinguish classes).

AUC < 0.5: The model is worse than the random guess (as if the predictions were inverted).

- Matthews Correlation Coefficient (MCC)

MCC is a metric that evaluates the performance of binary classification models and aims to measure the accuracy of the model in a more comprehensive way. MCC calculates the overall accuracy of the classification model by considering TP, FP, TN and FN values. MCC is a powerful metric for assessing the overall accuracy of the model and is particularly useful in unbalanced data sets. this metric, which is easily available through Weka, provides a more accurate understanding of model performance.

3. Results and Discussion

The dataset was analyzed with 11 different methods. In the first stage, since the results obtained with the random forest1 and simple logistic function1 methods were higher than the others (96.34% and 96.08% accuracy rates), these two methods were re-run. While the first results were obtained with cross validation (10), percentage split (80) was studied as second and in this case, the ACC value as random forest2 reached 97.4% and the ACC value as simple logistic function2 reached 96.1%. While there was an increase of 0.02% in the simple logistic function method, the difference increased to 1.06% in the random forest method. The value of 97.4% obtained with the random forest2 method is higher than the studies in the literature.

When the studies on thyroid are examined; Tantika achieved an ACC rate of 93.8% with Support Vector Machine (SVM) method in his conference presentation in 2022. The data set was 7200 samples [29]. Yusuf and Hidayatulloh reached 96.1% with Artificial Neural Network method in a data set of 383 samples in 2024 [30]. Angel and Herwindiatib achieved 97% with SVM, 92% with Radial Basis Function (RBF) and 91% with K-Nearest Neighbor (KNN) in 2024. The dataset consisted of 9172 samples [31]. Wijonarko achieved ACC rates ranging from 91.803% to 98.491% with Navie Bayes method on a dataset of 265 samples in 2018 [32]. Faruqziddan et al. reached 97.5% with RF method on a 383-sample data set [33]. In his thesis study, Luthfi found 95.16% with RF, 87.10% with KNN, 93.55% with Naive Bayes and 90.32% with Decision Tree on a data set of 123 samples [34]. Mutawali et al. obtained 93% with KNN algorithm [35], Untuk and Harga obtained 64% with Naive Bayes method, 63% with KNN [36], Tamba obtained 82.60% with RF [37]. In these studies [29-37] (table 3), the highest ACC rate was seen as 97% [31] and the lowest ACC rate was seen as 63% [36]. The 97% rate was obtained with 9172 samples. This result was achieved with the SVM method. The 63% rate was obtained with the KNN method. The proposed study reached 97.4% ACC rate with the random forest method. This result is a higher value than all the results in the literature [29-37].

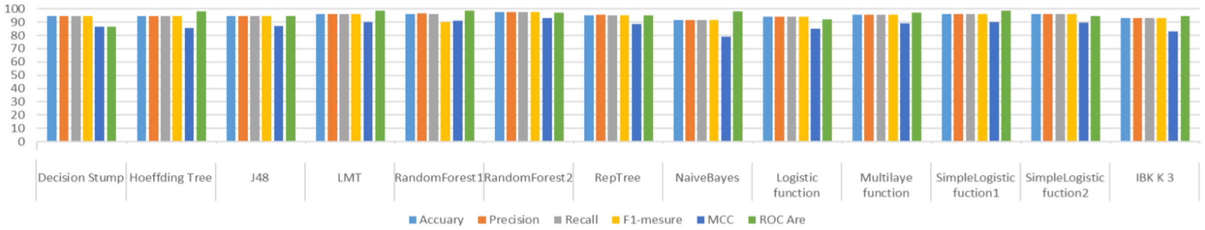


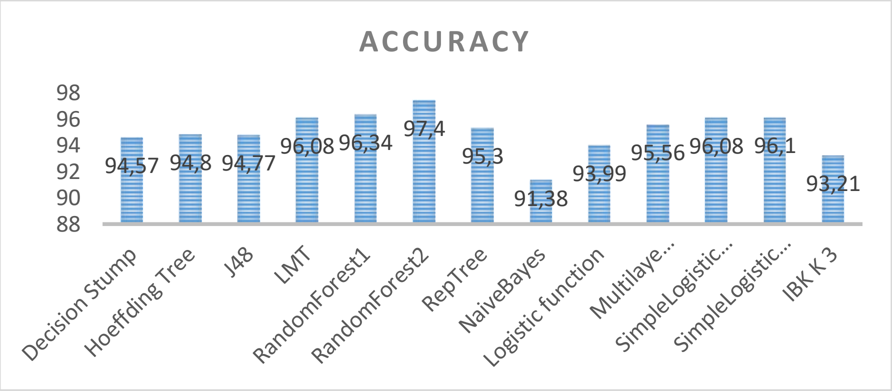
Fig. 2. Graphical representation of the results

Table 2. Results (ACC, precision, recall etc) of the methods (decision stump, hoeffding tree, J48 etc) studied

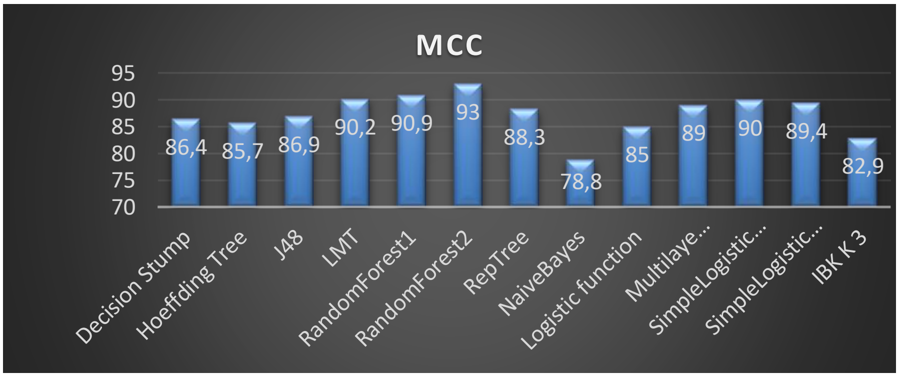
Model	Verification technique	Accuracy	Precision	Recall	F1-Mesure	Mcc	Roc are
Decision Stump	Cross Validation (10)	94.57	94.7	94.5	94.4	86.4	86.4
Hoeffding Tree	Percentage Split (80)	94.8	94.8	94.8	94.7	85.7	98.3
J48	Cross Validation (10)	94.77	94.8	94.8	94.7	86.9	94.7
LMT	Cross Validation (10)	96.08	96.1	96.1	96	90.2	98.7

Random Forest1	Cross Validation (10)	96.34	96.4	96.3	90.3	90.9	98.8
Random Forest2	Percentage Split (80)	97.4	97.4	97.4	97.4	93	97.2
Rep Tree	Cross Validation (10)	95.3	95.4	95.3	95.2	88.3	95.2
Naive Bayes	Cross Validation (10)	91.38	91.4	91.4	91.4	78.8	97.9
Logistic Function	Cross Validation (10)	93.99	93.9	94	93.9	85	92.2
Multilayer Perception Function	Cross Validation (10)	95.56	95.5	95.6	95.5	89	97.3
Simple Logistic Function1	Cross Validation (10)	96.08	96.1	96.1	96	90	98.7
Simple Logistic Function2	Percentage Split (80)	96.1	96.1	96.1	96.1	89.4	94.5
IBK K 3	Cross Validation (10)	93.21	93.2	93.2	93.1	82.9	94.4

When the dataset was studied with the cross validation (10) technique with the method we call random forest1, it gave the best results among other methods with 96.34 accuracy, 96.4 precision, 96.3 recall and 98.8 ROC ARE ratios. When percentage split (80) validation technique was used with the method we call random forest2, accuracy, precision, recall and f1-duration rates increased to 97.4% and MCC increased to 93%, reaching the best results among other methods. It was only behind a few methods in terms of ROC-ARE ratio (figure 2, figure 3.a, figure b and table 2). According to the literature review, both random forest1 and random forest2 yielded higher results than the studies (table 3).



(a)



(b)

Fig. 3. (a) Accuracy values of the studied methods, (b) MCC values of the studied methods

Table 3. Comparison of the proposed method with the literature

Authors	Year	Method	Results (ACC)
Tantika (29)	2022	SVM	%93,8
Yusuf ve Hidayatulloh (30)	2024	ANN	%96,1
Angel ve Herwindiatib (31)	2024	SVM	%97
Angel ve Herwindiatib (31)	2024	RBF	%92
Wijonarko (32)	2018	Navie Bayes	%91,803-%98,491
Faruqziddan v.d. (33)	2024	RF	%95,16
Luthfi (34)	2023	RF	%95,16
Luthfi (34)	2023	KNN	%87,10
Luthfi (34)	2023	Naive Bayes	%93,55
Luthfi (34)	2023	Decision Tree	%90,32
Mutawali [35]	2022	KNN	%93
Untuk ve Harga [36]	2020	Naive Bayes	%64
Untuk ve Harga [36]	2020	KNN	%63
Tamba [37]	2022	RF	%82,60
Recommended Method		Random Forest 2	%97,4

4. Conclusion and Suggestion

The 97.4% rate in the proposed method shows that the dataset analyzed in terms of thyroid cancer and recurrence status is an acceptable dataset, and it shows that it produces good results with the method we call random forest2. However, when 2.6% ratio is considered as a misdiagnosis as a result, considering an error that can lead to the death of people, studies to improve this will continue. The 97.4% ACC rate obtained is higher than the rates in the literature. For the 2.6% error rate found to be incorrect, specialist doctors can benefit from other imaging methods (ultrasound, tomography, MRI, etc.). This study can be used as a supporting mechanism for specialist doctors. It is thought that higher ACC rates can be achieved by obtaining more data from the patient (cholesterol, anamnesis, glucose, HbA1c etc.) and increasing the number of patients. Our future studies will also be aimed at reducing this incorrectly detected rate.

Acknowledgements

There is no conflict of interest with any person/institution in the prepared article. This study did not receive any specific funding or financial assistance from governmental, commercial, or nonprofit organizations.

References

[1] S. Ceylan, “Temel veri madenciliği algoritmalarının başarımlarının endokrin veri seti üzerinde karşılaştırılması,” Yüksek Lisans Tezi, Pamukkale Üniversitesi, Fen Bilimleri Enstitüsü, Elektrik-Elektronik Mühendisliği Anabilim Dalı, Turkey, 2023.

[2] G. Silahtaroğlu, *Veri madenciliği*. Turkey: Papatya Yayınları, 2008.

[3] S. Dogan and I. Turkoglu, "Extraction association rules from the biochemistry parameters for diagnosing hyperthyroidi," in *IEEE 16th SIU.*, (Aydin, Turkey), 2008, pp. 1-4, doi: 10.1109/SIU.2008.4632562.

[4] D. A. Koutras, “Subclinical Hyperthyroidism,” *Thyroid*, Vol. 9, No:3, pp: 311-315, 1999, doi: https://doi.org/10.1089/thy.1999.9.311

[5]“Hypertiroidism (overactive thyroid),” mayo.clinic.org. https://www.mayoclinic.org/diseases-conditions/hyperthyroidism/diagnosis-treatment/drc-20373665. (accessed Feb. 2024).

[6] A. Ozel, A. Akdemir and S. Orsel, “Hipertiroidinin Neden Olduğu Psikotik Bozukluk: Bir Olgu Sunumu,” *Nöropsikiyatri Arşivi*, 39(2-3-4), pp: 64-66, 2002. Available: https://www.researchgate.net/profile/Asena-

Akdemir/publication/242681514_Hipertiroidinin_Neden_Olduu_Psikotik_Bozukluk_Bir_Olgu_Sunumu/links/00b7d53989fe82f6e3000000/Hipertiroidinin-Neden-Olduu-Psikotik-Bozukluk-Bir-Olgu-Sunumu.pdf

[7] J. Barrera and R. M. Cesar-Jr, "An environment for knowledge discovery in biology," *Computers in Biology and Medicine,* 34, pp: 427–447, 2003, doi: https://doi.org/10.1016/S0010-4825(03)00073-8

[8] R. J. Shebuski, "Utility of point-of-care diagnostic testing in patients with chest pain and suspected acute myocardial infarction," *Current Opinion in Pharmacology,* 2, pp: 160–164, 2002, doi: https://doi.org/10.1016/S1471-4892(02)00140-6

[9] Ş. Kitiş and H. Göker, "Detection of obesity stages using machine learning algorithms," *Anbar Journal of Engineering Sciences*, 14(1), pp: 80-88, 2023. Available: https://www.iraqoaj.net/iasj/article/271320

[10] M. M. Yin and J. T. L.Wang, "GeneScout: a data mining system for predicting vertebrate genes in genomic DNA sequences," *Information Sciences,* 163, pp: 201–218, 2003, doi: https://doi.org/10.1016/j.ins.2003.03.016

[11] H. Göker, "Automatic detection of Parkinson's disease from power spectral density of electroencephalography (EEG) signals using deep learning model," *Physical and Engineering Sciences In Medicine*, vol.46, no.3, pp.1163-1174, 2023, doi: https://doi.org/10.1007/s13246-023-01284-x

[12] J. M. Ayub, et all. "Protein–Protein interaction map of the trypanosoma cruzi ribosomal p, protein complex," *Gene,* 357, pp: 129 – 136, 2005, doi: https://doi.org/10.1016/j.gene.2005.06.006

[13] Collaborators: jaina (Owner), "Thyroid Disease Data," kaggle.com. https://www.kaggle.com/datasets/jainaru/thyroid-disease-data/data (Accessed Feb. 2024).

[14] C. Boukhatem, H.Y. Youssef and A. B. Nassıf, "Heart disease prediction using machine learning," in *Advances in Science and Engineering Technology International Conferences* (United Kingdom), 2022, pp:1-6, doi: 10.1109/ASET53988.2022.9734880.

[15] I. Rish, "An emprical study of the naive bayes," IBM Research Report, USA, 2 November 2001.

[16] H. Zhang, "The optimality of naive bayes," in *FLAIRS Conference* (Miami Beach, Florida, USA), 2004, pp:1-6, Available: https://cdn.aaai.org/FLAIRS/2004/Flairs04-097.pdf

[17] E. Aydoğan, "Veri madenciliğinde sınıflandırma problemleri için evrimsel algoritma tabanlı yeni bir yaklaşım: rough-mep algoritması," Doktora Tezi, Gazi Üniversitesi, Turkey, 2008.

[18] S. Singaravelan, D. Murugan and R. Mayakrishnan, "Analysis of classification algorithms J48 and Smo on different datasets," *World Engineering & Applied Sciences Journal,* 6(2), 119-123, 2015, doi:10.5829/idosi.weasj.2015.6.2.22162

[19] S. Aljawarneh, M. B. Yassein and M. Aljundi, "An enhanced J48 classification algorithm for the anomaly intrusion detection systems," *Cluster Computing,* pp: 1-17, 2017, doi: https://doi.org/10.1007/s10586-017-1109-8

[20] M. A. Alan and C. Yeşilyurt, "Farklı veri setleri üzerinde SMO ve J48 algoritmalarının sınıflandırma sonuçlarının karşılaştırılması," *İşletme Bilimi Dergisi (JOBS),* 6(3), pp. 199-213, 2018, doi:10.22139/jobs.487388

[21] N. L. Leech, K.C. Barrett, and G.A. Morgan, *SPSS for intermediate statistics: Use and interpretation*. Manwah New Jersey, USA: Lawrance Erlbaum Associates Publishers, 2004.

[22] H. Bircan, "Lojistik regresyon analizi: Tıp verileri üzerine bir uygulama," *Kocaeli Üniversitesi Sosyal Bilimler Dergisi,* (8), pp: 185-208, 2004. Available: https://dergipark.org.tr/en/pub/kosbed/issue/25712/271314

[23] H. Tatlıdil, *Uygulamalı çok değişkenli istatistiksel analiz.* Ankara, Turkey: Akademi Matbaası, 1996.

[24] D. Delen, G. Walker and A. Kadam, "Predicting breast cancer survivability: a comparison of three data mining methods," *Artificial Intelligence in Medicine*, Vol 34, pp. 113-127, June 2005, doi: https://doi.org/10.1016/j.artmed.2004.07.002

[25] W. Chen and S. Zhang, "GIS-based comparative study of Bayes network, hoeffding tree and logistic model tree for landslide susceptibility modeling," *Catena*, 105344, 203, 2021, doi: https://doi.org/10.1016/j.catena.2021.105344

[26] N. Nahar and F. Ara, "Liver disease prediction by using different decision tree techniques," *International Journal of Data Mining & Knowledge Management Process (IJDKP),* Vol.8, No.2, March 2018, doi: 10.5121/ijdkp.2018.8201 1

[27] S. Ray, "A quick review of machine learning algorithms," in *2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon),* (India), 2019, pp. 35-39, doi: 10.1109/COMITCon.2019.8862451

[28] M. Zemouli, "Un système intelligent pour améliorer la prédiction des maladies cardiovasculaires," *Guelma: Université du*, 8 mai 1945 Guelma, 2023. Available: http://dspace.univ-guelma.dz/jspui/handle/123456789/15053

[29] B. Wijonarko, "Perbandingan algoritma data mining naive bayes dan bayes network untuk mengidentifikasi penyakit tiroid," *Jurnal Pilar Nusa Mandiri*, Vol. 14, No. 1, Maret 2018, doi: https://doi.org/10.33480/pilar.v14i1.83

[30] R. S. Tantika, “Penggunaan metode support vector machine klasifikasi multiclass pada data pasien penyakit tiroid,” *Bandung Conference Series: Statistics*, Vol. 2, No. 2, pp. 159-166, 2022, doi: https://doi.org/10.29313/bcss.v2i2.3590

[31] L. Yusuf and T. Hidayatulloh, “Implementasi algoritma artificial neural network dengan aktivasi ReLU: klasifikasi tiroid,” *Jurnal Swabumi*, Vol.12 No.1, pp. 113-119, Maret 2024. Available: https://repository.nusamandiri.ac.id/repo/files/248978/download/23020-59820-1-PB.pdf

[32] A. Angel and D. E. Herwindiatib, “Perbandingan algoritma K-NN, SVM, dan decision tree dalam klasifikasi kelenjar tiroid,” *Jurnal Teknologi Dan Sistem Informasi Bisnis*, Vol. 6 No. 4, hal. 866-871, Oktober 2024, doi: https://doi.org/10.47233/jteksis.v6i4.1651

[33] M. Faruqziddan, E. H. S. Aulia, S. D. Azzahra, A. Ristyawan and E. Daniati, “Klasifikasi risiko kambuhnya kanker tiroid menggunakan algoritma random forest,” *INOTEK*, Vol. 8, No. 1, hal. 63-74, Agustus 2024. Available: https://proceeding.unpkediri.ac.id/index.php/inotek/article/view/4912

[34] M. Luthfi, P. Kinerja “Algoritma klasifikasi untuk prediksi penyakit tiroid,” *Universitas Pembangunan Nasional “Veteran” Jakarta Fakultas Ilmu Komputer Program Studi Informatika*, Vol. 22, No. 2, 2023, doi: https://doi.org/10.31294/p.v21i2

[35] L. Mutawali, W. Murniati and K. Kunci, “Penerapan knnimputer dalam mengolah data missing value untuk membantu meningkatkan akurasi support vector machine klasifikasi penyakit tiroid,” 4.4 (2022): 386-390. 2022. Available: https://archive.ics.uci.edu/ml/datasets/thyroid+diseas

[36] C. Untuk and K. Harga, “Perbandingan kinerja algoritma decision tree dan naive bayes dalam prediksi kebangkrutan,” *Core*, vol. 7, no. 1, pp. 20–24, 2020. Available: https://core.ac.uk/download/pdf/143964255.pdf

[37] P. Tamba, “Prediksi penyakit gagal jantung dengan menggunakan random forest,” *Jurnal Sistem Informasi dan Ilmu Komputer Prima*, vol. 5, no. 2, 2022, doi: https://doi.org/10.34012/jurnalsisteminformasidanilmukomputer.v5i2.2445