Classification of Thyroid Disease by Using Data Mining Models: A Comparison of Decision Tree Algorithms

Ebru Turanoglu-Bekar¹, Gozde Ulutagay¹*, Suzan Kantarcı-Savas²

(1) Department of Industrial Engineering, Izmir University, Gursel Aksel Blv 14 Izmir, 35350, Turkey
(2) Department of Econometrics, Kirklareli University, Kayali Campus, Kofcaz, Kirklareli, 39700, Turkey

Copyright 2016 © Ebru Turanoglu-Bekar, Gozde Ulutagay and Suzan Kantarcı-Savas. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract
The main objective of this study is to provide a compact source of reference for the researchers who want to use decision tree which is an important tool of data mining technology in their area of work. With this aim in mind, we compared widely used decision tree algorithms to classify types of thyroid disease and compared their performances according to six performance metrics (ACC(%), MAE, PRE, REC, FME, and Kappa Statistic). We hope that this study can provide a useful overview of the current work of this field and highlight how to apply decision tree algorithms as a tool of data mining technology.

Keywords: Classification, Data Mining, Decision Tree, Thyroid Disease Diagnosis.

1 Introduction
Thyroid hormones created by the thyroid gland introduce two active thyroid hormones including levothyroxine (abbreviated T4) and triiodothyronine (abbreviated T3) to control the body’s metabolism. These hormones are necessary to help each cell in each tissue and organ to work right and to produce proteins in the regulation of body temperature, and in overall energy production and regulation [1-2]. Thyroid function affects every essential organ in the body. The significance of thyroid disorders should not do justice to thyroid storm (an episode of severe hyperthyroidism) and myxedema coma (the end stage of untreated hypothyroidism) may cause death in a substantial number of cases [3].
The disease of thyroid is originally categorized as either functional or structural. The functional classification that is initial biological investigation of thyroid disorders includes euthyroid (normal), hyperthyroid (overactive) and hypothyroid (underactive) shown as in Figure 1.

The structural categorizations are performed consistent with the morphology of the gland defined either by palpation (physical examination) or by some visualizing methods such as ultrasonography (USG) that utilizes sound waves or scintigraphy which makes use of radioactive materials [4].

![Figure 1: The functional classification of thyroid disorders according to TSH level](image)

The functional state of the thyroid is essential and it denotes the basis for diagnosis and therapy in most thyroid diseases. According to Figure 1, Euthyroidism implies standard production of thyroid hormones by the thyroid and standard levels in the circulation and at the cellular stage [5-6]. Hyperthyroidism is defined clinical symptomatology resulting from overmuch circulating and intracellular thyroid hormones and Hypothyroidism occurs almost always because of nonexistence thyroid hormone production and insufficient replacement therapy [7-8].

The classification of thyroid diseases is regarding the identification of new disease existences, better understanding of the molecular and immune systmes accountable for failings, and disease evolution. American Thyroid Association presents that the classification “… had to be surveyed intermittently and reviewed as additional information might require.” Accordingly, the precise diagnosis of thyroid dysfunctions established on clinical and laboratory tests is a significant work [9]. Laboratory tests on serum examples have to be employed to define the existence of hyperthyroidism or hypothyroidism. But, it is difficult to determine on the structural classification of the disease with these tests. For example, a patient categorized as deceiving hyperthyroidism with regard to the serum hormone level may possess a thyroid gland of nodular form or diffuse form and the structural characteristic decides the treatment protocol.
Thyroid dysfunction diagnosis is a significant classification problem and it also demonstrates a difficulty to
diagnose by traditional parametric and nonparametric statistical methods such as discriminant analysis,
logistic regression, k-nearest-neighbor, mathematical modelling and etc. [3] since they perform well only
when the basic assumptions are satisfied. The effectiveness of these methods relies on a great amount on the
various assumptions or conditions under which models are established. Researches must know both data
properties and model capabilities before they can effectively apply the model. Consequently, recent
improvements in diagnosis of diseases indicate that diagnostic expert systems can support to researchers to
make a diagnosis effectively.

Therefore, the main objective of this study is to compare the well-known decision tree algorithms such as J
48, CART, NBT, BFTree, LADTree, REPTree, RANDOMTree, RANDOM FOREST, LMT, FT, and DS
in diagnosing thyroid disease. These algorithms are performed on thyroid data set created by a specialist.
The thyroid data set analyzed with a different viewpoint by using different types of decision tree algorithms
used on previous researches. Also, their performances are comprised according to six performance metrics
(ACC (%), MAE, PRE, REC, FME, and Kappa Statistic). This paper is structured as follows: the following
second section gives the necessary background information about the decision trees classification methods.
The third section explains the decision trees algorithms used in this study. The fourth section is reserved for
explanation of the experiment and discussion. Finally, the last section is the conclusion part.

2 Background

2.1. Literature review on diagnosis of thyroid diseases
A search for literature on diagnosis of thyroid diseases indicates that different data mining and artificial
intelligence methods such as fuzzy logic, artificial neural networks, fuzzy neural networks, support vector
machine, immune recognition system, etc. have been used to determine the type of thyroid given as in Table
1.

| References           | Publication journal                  | Used Method/Methods                                                   |
|----------------------|--------------------------------------|                                                                      |
| Sharpe et al. [10]   | Clinical Chemistry                   | Artificial Neural Networks (ANN)                                     |
| Serpen et al. [11]   | In Proceedings of artificial neural networks in engineering conference | Multi-Layer Perceptron (MLP) Learning Vector Quantizer (LVQ) Radial Basis Function (RBF) Probabilistic Potential Function Neural Network (PPFNN) |
| Bramejeer and Banzhaf [12] | IEEE Transactions on Evolutionary Computation | Linear genetic programming (GP)                                      |
| Zhang and Berardi [3] | Health Care Management Science       | ANN                                                                  |
| Özyılmaz and Yıldırım [13] | In Proceedings of ICONIP’02 ninth international conference on neural information processing | MLP with Back-Propagation MLP with Fast Back- Propagation RBF Adaptive Conic Section Function Neural Network (CSFNN) |
| Pasi [14]            | In International conference on soft computing | Linear Discriminant Analysis (LDA) C4.5 MLP DIMLP with two hidden layers and default learning parameters (DIMLP) |
| Hoshi et al. [15]    | Chemical and Pharmaceutical Bulletin | Self-organizing map (SOM) Bayesian regularized neural network (BRNN) |
| Authors              | Journal/Media                          | Techniques/Models                                      |
|---------------------|----------------------------------------|--------------------------------------------------------|
| Polat et al. [16]   | Expert System with Application         | The artificial immune recognition system (AIRS)        |
|                     |                                        | AIRS with fuzzy weighted pre-processing                |
| Erol et al. [4]     | Journal of Medical Systems             | MLP                                                    |
|                     |                                        | RBF                                                    |
| Keles and Keles [17]| Expert System with Application         | Neuro Fuzzy Classification                             |
| Temurtas [1]        | Expert System with Application         | Multilayer neural network                              |
|                     |                                        | Probabilistic Neural Network (PNN)                     |
|                     |                                        | LVQ                                                    |
| Khanale and Ambilwade [18] | Journal of Artificial Intelligence | Fuzzy Inference System (FIS)                           |
| Dogantekin et al. [19] | Expert System with Application      | Wavelet Support Vector Machines (WSVM)                |
|                     |                                        | Generalized Discriminant Analysis (GDA)                |
| Chen et al. [20]    | Journal of Medical Systems             | Support Vector Machines (SVM)                          |
|                     |                                        | Particle Swarm Optimization (PSO)                      |
|                     |                                        | Grid Search Technique                                  |
| Jaganathan and Rajkumar [21] | International Journal of Computational Science and Engineering | MLP                                              |
|                     |                                        | GDA                                                    |
|                     |                                        | WSVM                                                   |
| Li et al. [22]      | Journal of Medical Systems             | Computer Aided Diagnosis (CAD)                         |
|                     |                                        | Principle Component Analysis (PCA)                     |
|                     |                                        | Extreme Learning Machine (ELM)                         |
| Liu et al. [23]     | Journal of Medical Systems             | Fuzzy K-Nearest Neighbor (FKNN) Classifier             |
|                     |                                        | PSO                                                    |
|                     |                                        | PCA                                                    |
| Azar et al. [24]    | Communications in Computer and Information Science | Linguistic Hedges Neural-Fuzzy Classifier with Selected Features (LHNFCSF) |
| Ulutagay [2]        | Wulfenia Journal                       | FIS                                                    |
| Rawte and Roy [25]  | International Journal of Engineering Research & Technology | Ontology Based Expert System                          |
| Maysanjaya et al. [26] | In Proceedings of International Seminar on Intelligent Technology and Its Applications, (ISITIA 2015) | MLP with Back-Propagation                            |
| Biyouki et al. [27] | IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology, (CIBCB 2015) | Fuzzy Rules by k-means Algorithm                      |
|                     |                                        | Scaled Conjugate Gradient Algorithm (SCG)              |
| Prasad et al. [28]  | Soft Computing                         | String Matching System                                |
|                     |                                        | Artificial Bee Colony Optimization                    |
|                     |                                        | PSA                                                    |
|                     |                                        | Rough Data Sets Theory                                 |

Based on the results of literature review, very little progress has been used by the decision trees to diagnose of thyroid diseases. In literature, Lavanya and Usha Rani [19] investigated the performance of decision tree classifiers on different medical data sets. The performance analysis indicated that CART has the highest performance for classification of medical data for this data set. Also, Margret et al. [20] paid attention into the diagnosis of thyroid disorders using decision tree attribute splitting rules. As a result of this study, the thyroid data set classified into three classes of thyroid disorders.

### 2.2. Decision tree

Decision Trees are the most popular architectures widely used in data mining [31]. These architectures use a divide-and-conquer strategy in order to partition the instance space into decision regions. At first, a root node is designated by using a test. Then, the value of related test attribute splits the data set and, the process
is repeated until the determined stopping criterion is provided. At the end of the tree, each node is named as leaf node. Each leaf node denotes the class. Also, each branch indicates a path defined as a decision rule. The classification is handled by using each decision rule for a new sample [32]. As a summary, the decision tree architectures consist of a root node, branches, internal nodes, and leaf nodes. There are three main steps for classification by using decision trees: The first step is the learning process. The model is constructed on the training data. Hence, this model is presented by classification rules. In the second step, a test is selected in order to calculate the model accuracy. The model is accepted according to the value of this test. If this value is considerably accepted, the model could be used for the classification of a new datum. At last, the third step includes the usage of the model for a classification or prediction of a new data (Figure 2).

![Figure 2: Steps for Classification and Prediction Process](image)

There are different kinds of algorithms for the construction of decision trees such as CART, C4.5, etc [33-35]. The differences among the algorithms are formed due to the selected criteria for the test attribute. In subsections, some basic information is given about the decision tree algorithms used in the study.

3 Decision tree algorithm

3.1. J 48

J 48 is modified version of C 4.5 [34]. The principle of this algorithm is to use divide-and-conquer strategy. Also, it uses pruning for the construction of the tree in order to avoid over-fitting problem. Maximum gain information is used as a splitting criterion. It calculates overall entropy of the training data and entropy for each attribute according to classes. Then it takes the differences between overall entropy and entropy achieved for each attribute. This value is called as the gain information. Then, the attribute which has the highest gain information is selected for splitting. The formulas for entropy and gain information are given below, respectively:

\[
\text{Entropy}(S) = -\sum_{j=1}^{k} \frac{\text{freq}(C_j,S)}{|S|} \log_2 \frac{\text{freq}(C_j,S)}{|S|}
\]

\[
\text{Gain}(S,A) = \text{Entropy}(S) - \sum_{i=1}^{k} \frac{|S_i|}{|S|} \text{Entropy}(S_i)
\]

where \(S\) is a training set explained in terms of \(k\) attributes, and \(C = \{C_1,C_2,\ldots,C_n\}\) defines \(n\) classes.

3.2. CART

Classification and Regression Trees (CART) is a kind of statistical technique [33]. It is a tree structure technique that produces binary trees. Each internal nodes has two outgoing edges according to the selected test attribute. There are various types of impurity criteria by using CART such as Gini Index, Symmetric Gini Index, Twoing, Ordered Twoing, Class Probability for Classification Tree, Least Squares, Least Absolute Deviation for Regression Trees, Multi Variable Splitting Criterion and The Linear Combinations Method. For example, Gini Index is an impurity-based criterion. The pruning of the constructed tree is done
by cost–complexity pruning. The divergences among the probability distributions of the target attribute’s value are measured by cost-complexity pruning. It is given as follows [36]:

\[
Gini(y,S) = 1 - \sum_{c_j \in \text{cat}(y)} \left( \frac{\sigma_{y=c_j}|S|}{|S|} \right)^2
\]

(3.3)

3.3. NBTree

NBTree combines Naive Bayes Classification and Decision Trees [37]. The decision tree constructed by NBTree algorithm uses Naive Bayes Classifiers. The tree contains univariate splits. This algorithm uses Bayes rule in order to find the probability of each class given the instance. This algorithm assumes that the attributes are conditionally independent given the label. NBTree classifier generally has higher accuracy rate than a naïve Bayes classifier.

3.4. BFTree

Best-First decision tree (BFTree) learning process uses the procedure defined for standard decision trees [38]. It handles categorical and numerical variables. While standard decision tree induction process expands in depth-first order, the best-first decision tree induction process expands the “best” node first.

3.5. LADTree

LADTree is a classification method. In this approach, decision trees are combined with the predictive accuracy of boosting into a set of classification rules [39]. Then, they are adapted to the multiclass LogitBoost and AdaBoost. LogitBoost algorithm is fused with AdaBoost induction. Multi-class alternating decision trees are built by using this structure. A single variable is selected for the splitter node. The algorithm aims to minimize the least squares between the working return and the mean value of the examples.

3.6. REPTree

REPTree is decision tree learner which builds a decision and regression tree [35]. The gain information is used in order to split the data set. It prunes the tree in order to reduce error pruning. The main purpose is not only to decrease the effectiveness of the noisy instances but also to decrease the complication in the classification progress. REPTree is a fast algorithm.

3.7. RANDOMTree

RANDOMTree is a multiple random tree algorithm [40, 41]. A non-tested attribute is selected randomly from the whole data set without using a training set. A limit is predefined. The tree has been constructed until the depth of the tree exceeds this predefined limit. If the depth of the tree exceeds this limit, it stops. The training set is used in order to update the statistics of each node. The class attribute contains different classes. Each node sets down the number of records classified as different classes. The same process is applied for the classification by using a decision tree. Each data record is read in order to update multiple random trees. It is necessary to complete one scan of the data. The classification of a datum x is performed by averaging the probability outputs from multiple random trees.

3.8. Random forest

Random Forest is a group of tree predictors [42]. A random vector is used. It is sampled independently by using the same distribution \( \theta_k \) is handled from the old vectors \( \theta_1, \theta_2, \ldots, \theta_{k-1} \). X is defined as an input vector. The construction of the tree is handled on the training set by using the random vector \( \theta_k \). The resulting is defined with \( h(X, \theta_k) \). If a large numbers of trees are generated, they are voted in order to find the most
popular class. The procedure is called as random forests. It is a classifier. Each tree has a cost as a vote for the class selected the most popular at input X.

3.9. LMT

Logistic Model Trees (LMT) associates tree induction and logistic regression technique [43]. It constructs a single tree. It contains binary splits on numeric attributes, multi-way splits on nominal ones, and logistic regression models at the leaves. The logistic regression states at each parent node and a leaf node gathers all parent models for estimating a probability for each classes. Also, a trimming process is used to generalize the model.

3.10. FT

Functional Trees (FT) is a kind of multivariate tree [44]. They are able to explore multiple representation way by using decision tests. This algorithm works as a multivariate tree learning algorithm. Decision nodes with multivariate tests, leaf nodes made predictions using linear functions are used for generating the functional trees. This algorithm can handle missing values. It can be applied into binary, numerical or categorical data.

3.11. Decision stump

Decision Stump (DS) is a single level decision tree [45]. It is a kind of machine learning model. The decision tree constructed by DS consists of one root node and this node connects to the leaf nodes. Just a single input attribute is used in order to make a prediction. It is regularly employed with a bagging or boosting algorithm.

4. Experimental study

The main purpose is to search the best classification approach for thyroid disease diagnosis by making the comparison of decision tree algorithms. In the line of this purpose, the experiments are conducted to compare different kinds of decision tree algorithms given in the previous section. The steps used in the study are stated as the following subsections.

4.1. Normalization

In this study, min-max normalization is performed as a data transformation technique. It is a kind of linear transformation. Let A is an attribute, $\text{min}_A$ and $\text{max}_A$ are the minimum and the maximum values of this attribute. This normalization technique maps a value $v$ of A into $v'$ in a new range $[\text{newmin}_A, \text{newmax}_A]$ by using the following formula [46,47]:

$$v' = \frac{v - \text{min}_A}{\text{max}_A - \text{min}_A} \cdot (\text{newmax}_A - \text{newmin}_A) + \text{newmin}_A$$

In this case, min-max normalization maps a value $v$ of A into $v'$ in a new range between 0 and 1. The following simplified formula is used for min-max normalization:

$$v' = \frac{v - \text{min}_A}{\text{max}_A - \text{min}_A}$$

4.2. Training and testing process

CART, NBTree, BFTree, ADTree, REPTree, Random Tree, Random Forest, LMT, FT, DS algorithms were used for the experiment. The WEKA framework was used in order to implement these algorithms. The computational experiments have been performed on a system with 1.60GHz Core (TM) i5, 4GB RAM and running on Windows 7.

In the training and testing process, 10-fold cross validation was performed on the dataset. Five performance metrics were handled for comprising the algorithms given below:
Classification Accuracy (ACC) is a widely used measure to evaluate a classifier. It is just defined as the degree of right predictions of a classifier. It can take values range from 0 to 100(%).

\[
ACC = \frac{\text{num(test examples rightly classified)}}{\text{num(total test examples)}}
\]  

(4.6)

Precision (PRE) is a kind of measure. It assures a specific class which has been forecasted. It can be thought as percentage of times that the classifier is correct in its classification of positive samples.

\[
PRE = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalsePositive}}
\]  

(4.7)

Recall (REC) measure the capability of a prediction model for selecting the samples from the same class.

\[
\text{REC} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalseNegative}}
\]  

(4.8)

Whereas a true positive can be defined as a positive sample identified with the same label correctly, a false positive can be defined as a negative sample identified incorrectly with the positive label. Also, a true negative means that a negative sample identified with the same label correctly. On the other hand, a false negative is a positive sample identified with the negative label incorrectly.

\[F\text{-Measure (FME)}\] is the harmonic mean of precision and recall.

\[
FME = \frac{2 \times \text{PRE} \times \text{REC}}{\text{PRE} + \text{REC}}
\]  

(4.9)

Mean Absolute Error (MAE) can be explained as the average of the absolute values of the prediction errors. It demonstrates the deviations from the true probability by calculating the absolute value of differences.

\[
MAE = \frac{\sum_{j=1}^{c} \sum_{i=1}^{m} |p(i, j) - p(i, j)|}{mxc}
\]  

(4.10)

Kappa Statistic is a measure of agreement between two categorical variables, X and Y. It is used to make comparison for the ability of different raters. The Kappa Statistic takes values between 0 and 1. The observed agreement between X and Y is shown by p_0 and the expected agreement by chance is shown by p_e and the kappa statistic is defined as follows:

\[
\hat{K} = \frac{p_0 - p_e}{1 - p_e}
\]  

(4.11)

Confidence \( c \) of a rule \( A \Rightarrow B \) is a measure of the accuracy rule calculated as follows [31]:

\[
\text{Confidence} = P(B/A) = \frac{P(A \cap B)}{P(A)}
\]  

(4.12)

4.3. Results and discussion

The comparison of the decision tree algorithms performed on Thyroid Data is shown in Table 2. NBTree has the highest accuracy rate. Accuracy rate of this algorithm is 75%. It is seen that NBTree is the most powerful classifier for this sample. This result shows that the thyroid disease of a new patient is predicted successfully with an acceptable ratio 75%. J48 and LADTree has the same accuracy rate as 66.25%.
If Precision (PRE) value is close to 1, it can be concluded that there is a perfect precision. NBTree has the closest value to 1. The precision value was found as 0.773 which indicates that the most of the samples predicted as belonging to one class do truly belong to the same class.

If Recall (REC) value is approximately 1, perfect recall is handled. In this experiment, NBTree has the highest REC value as 0.75. It can be explained as most classification items has been classified in own class.

MSE is also an important measure among the performance metrics. If it is approximately 0, it shows that the difference between the observed and the predicted values is too small. Random Tree has the minimum MSE value (0.125) for this experiment. But the accuracy rate of this algorithm is under the average accuracy rate of all algorithms as 62.5. It is seen that next one is Random Forest with 0.126.

The kappa statistic is equal to 1 means a perfect agreement. NBTree has the closest value to 1. It can be interpreted as substantial agreement with 0.684 between the predictions with the true class. J48 and LADTree have the same kappa statistic value with 0.568. A moderate agreement is seen for these two algorithms. Decision Stump has the worst kappa statistic value with 0.239 which shows slight agreement.

Table 2: Comparing the various decision tree algorithms carried out thyroid data.

| Algorithm  | Tree Size | Leaf | ACC(%) | MAE   | PRE   | REC   | FME   | Kappa Statistic |
|------------|-----------|------|--------|-------|-------|-------|-------|-----------------|
| NBTree     | 5         | 3    | 75.00  | 0.1548| 0.773 | 0.75  | 0.749 | 0.684          |
| J48        | 21        | 11   | 66.25  | 0.1673| 0.581 | 0.638 | 0.603 | 0.568          |
| LADTree    | 21        | 14   | 66.25  | 0.1312| 0.668 | 0.663 | 0.663 | 0.568          |
| BFTree     | 17        | 9    | 65.00  | 0.1378| 0.666 | 0.65  | 0.651 | 0.556          |
| LMT        | 1         | 1    | 65.00  | 0.1797| 0.622 | 0.65  | 0.626 | 0.545          |
| Random Forest | 10 Trees |      | 65.00  | 0.1264| 0.665 | 0.65  | 0.646 | 0.551          |
| FT         | 5         | 3    | 63.75  | 0.1673| 0.581 | 0.638 | 0.603 | 0.529          |
| Random Tree | 51       |      | 62.50  | 0.125 | 0.627 | 0.625 | 0.625 | 0.523          |
| REPTree    | 13        |      | 62.50  | 0.1702| 0.634 | 0.625 | 0.625 | 0.523          |
| CART       | 21        | 11   | 58.75  | 0.1715| 0.577 | 0.588 | 0.573 | 0.470          |
| DS         | Single Level |      | 41.00  | 0.2288| 0.185 | 0.413 | 0.256 | 0.239          |

Also, it is seen that Decision Stump has the worst accuracy rate with 41%. This tree has one level. The precision and recall values are the worst values among the other algorithms values.

Friedman Two Way Analysis of Variances by Ranks Test is applied for the comparison of the algorithms. The aim of this test is to determine whether there is any difference among the algorithms effect or not. With four algorithms and six performance metrics (ACC (%), MAE, PRE, REC, FME, and Kappa Statistic), the test statistic distributed with $X^2_9$ is obtained as 29.401. p value is equal to 0.001. It is seen that p value is below 0.05 (p=0.001<0.05) alpha level significance. The null hypotheses were rejected. It is concluded that the algorithms have differences. The results of pairwise comparisons for eleven algorithms are showed in Table 3.

It is seen that the pairwise comparison between DS-LMT (p=0.019<0.05), DS-BFTree (p=0.012<0.05), DS-LADTree (p=0.002<0.05), DS-NBTree (p=0.000<0.05), CART-LMT (p=0.037<0.05), CART-BFTree (p=0.024<0.05), CART-LADTree (p=0.040<0.05), CART-NBTree (p=0.000<0.05), RandomTree-LADTree (p=0.012<0.05), RandomTree-NBTree (p=0.001<0.05), FT-LADTree (p=0.045<0.05), FT-NBTree (p=0.007<0.05), and REPTree-NBTree (p=0.012<0.05) are significant. 40% of these significant comparisons include NBTree.
Table 3: The results of pairwise comparisons for the algorithms using Friedman’s 2-way ANOVA.

| Algorithm1 vs. Algorithm2 | Test Statistic | Significance | Algorithm1 vs. Algorithm2 | Test Statistic | Significance |
|---------------------------|----------------|--------------|---------------------------|----------------|--------------|
| DS_CART vs. DS-REPTree    | 0.500          | 0.794        | FT-REPTree vs. FT-J48     | -0.333         | 0.862        |
| DS-REPTree vs. DS-J48     | 1.167          | 0.542        | FT-J48 vs. FT-LMT         | 1.500          | 0.433        |
| DS-J48 vs. DS-LMT         | 2.167          | 0.258        | FT-LMT vs. FT-BFTree      | 2.333          | 0.223        |
| DS-LMT vs. DS-BFTree      | 3.667          | 0.056        | FT-BFTree vs. FT-LADTree  | 2.667          | 0.164        |
| DS-BFTree vs. DS-LADTree  | 4.000          | 0.370        | FT-LADTree vs. REPTree-J48| 3.833          | 0.045        |
| DS-LADTree vs. DS-NBTree  | 6.000          | 0.002        | REPTree-J48 vs. J48-RandomForest| 1.500          | 0.433        |
| DS-NBTree vs. CART-RandomTree| 7.333        | 0.000        | REPTree-J48 vs. J48-LMT   | 2.000          | 0.296        |
| CART-RandomTree vs. CART-FT| 0.667         | 0.728        | REPTree-J48 vs. J48-LADTree | 2.333          | 0.223        |
| CART-FT vs. CART-REPTree  | 1.667          | 0.384        | REPTree-J48 vs. J48-NBTree | 3.500          | 0.068        |
| CART-REPTree vs. CART-J48 | 2.000          | 0.296        | REPTree-J48 vs. J48-LADTree | 3.500          | 0.068        |
| CART-J48 vs. CART-RandomForest| 3.167        | 0.098        | REPTree-J48 vs. J48-LADTree | 3.500          | 0.068        |
| CART-RandomForest vs. CART-LMT| 3.500         | 0.068        | REPTree-J48 vs. J48-LADTree | 3.500          | 0.068        |
| CART-LMT vs. CART-BFTree  | 4.000          | 0.037        | REPTree-J48 vs. J48-LADTree | 3.500          | 0.068        |
| CART-BFTree vs. CART-LADTree| 4.333         | 0.024        | REPTree-J48 vs. J48-LADTree | 3.500          | 0.068        |
| CART-LADTree vs. CART-NBTree| 5.500         | 0.000        | REPTree-J48 vs. J48-LADTree | 3.500          | 0.068        |
| CART-NBTree vs. Random Tree-FT| 6.833         | 0.000        | REPTree-J48 vs. J48-LADTree | 3.500          | 0.068        |
| Random Tree-FT vs. Random Tree-REPTree| 1.000         | 0.602        | REPTree-J48 vs. J48-LADTree | 3.500          | 0.068        |
| Random Tree-REPTree vs. Random Tree-J48| -1.333        | 0.486        | REPTree-J48 vs. J48-LADTree | 2.000          | 0.296        |
| Random Tree-J48 vs. Random Tree-LMT| 2.500         | 0.192        | REPTree-J48 vs. J48-LADTree | 3.500          | 0.068        |
| Random Tree-LMT vs. Random Tree-BFTree| 3.667         | 0.056        | REPTree-J48 vs. J48-LADTree | 3.500          | 0.068        |
| Random Tree-BFTree vs. Random Tree-LADTree| 4.833         | 0.012        | REPTree-J48 vs. J48-LADTree | 3.500          | 0.068        |
| Random Tree-LADTree vs. Random Tree-NBTree| 6.167         | 0.001        | REPTree-J48 vs. J48-LADTree | 3.500          | 0.068        |

The comparison of decision tree algorithms with respect to runtime is shown in Figure 3. REPTree, Random Tree, and Decision stump are faster than others. LMT algorithm takes a long time even though small data set is used. While J48 and LADTree has the same accuracy rate, J48 is faster than LADTree.
It is seen in Table 4 that Pituary Hypopituitarism and Hyperthyroid classes have the highest True Positive Rate with 75%. It means that 75% data defined by these classes are classified as a given class. Subclinical Hyperthyroid class has the minimum False Positive Rate with 1.4%. False Positive is used in order to define instances falsely classified as a given class. Yet, Subclinical Hyperthyroid class has True Positive Rate 50%. Precision is defined as the proportion of True Positive instances to the total of True Positive and False Positive instances. It is shown that Hyperthyroid class has the maximum precision value with 71%. Also, Recall is another proportion. It is calculated by using the proportion True Positives to the total of True Positive and False Negative instances. As it is seen before, Pituary Hypopituitarism and Hyperthyroid classes have the highest values. Their Recall value is 75%. F-Measure combines precision and recall. If F-Measure is used for the comparison of accuracy by class, Hyperthyroid has the highest value with 73%. The whole class has higher value than 50% for ROC Area.

Table 4: Detailed Accuracy by Class.

| Class            | TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area |
|------------------|---------|---------|-----------|--------|-----------|----------|
| Hypothyroid      | 0.625   | 0.161   | 0.625     | 0.625  | 0.625     | 0.619    |
| Euthyroid        | 0.700   | 0.117   | 0.667     | 0.700  | 0.683     | 0.816    |
| Pituary Hypopituitarism | 0.750   | 0.042   | 0.667     | 0.750  | 0.706     | 0.848    |
| Subclinical Hypothyroid | 0.500   | 0.026   | 0.500     | 0.500  | 0.500     | 0.785    |
| Hyperthyroid     | 0.750   | 0.078   | 0.706     | 0.750  | 0.727     | 0.771    |
| Subclinical Hyperthyroid | 0.500   | 0.014   | 0.800     | 0.500  | 0.615     | 0.853    |

In the literature, it is seen that generally C4.5 and CART algorithms are used on medical data set [16,18]. They have good enough performances on medical datasets. However, NBTree is the most powerful classifier in our case. On the other hand, J48 is more interpretable. It is the modified version of C4.5 in WEKA. The tree generated by J48 is given in Figure 3. This tree can be used to classify a new patient according to rules generated from this tree structure. The rules are given in Table 5. Also, it is seen that there are confidence values for each rule in this Table. The analyst can set a minimum confidence level of 60%. As a result, the rule 9 cannot be reported for the classification. It means that the rules which have higher confidence values than the minimum confidence level, they cannot be used in order to make the classification.

For example, suppose that a new patient has TSH value 0.98, FT3 value 1.8, and FT4 value 19.7. When we use the rule 10 given in Table 5, Thyroid Disease could be diagnosed for a new patient as Euthyroid.
### Table 5: Detailed accuracy by class.

| Rules generated for Pituitary Hypopituitarism | Confidence |
|---------------------------------------------|------------|
| 1 If TSH ≤ 0.91 and FT4 ≤ 12.26 then Pituitary Hypopituitarism | 8/10 = 0.80 |

| Rules generated for Subclinical Hypothyroid | Confidence |
|-------------------------------------------|------------|
| 2 If TSH > 0.91 and FT3 > 0.80 and TSH > 5.41 and FT3 ≤ 1.34 then Subclinical Hypothyroid | 3/3 = 1.0 |

| Rules generated for Hypothyroid | Confidence |
|---------------------------------|------------|
| 3 If TSH > 0.91 and FT3 ≤ 0.80 and TSH > 2.26 then Hypothyroid | 15/15 = 1.0 |
| 4 If TSH > 0.91 and FT3 ≤ 0.80 and TSH ≤ 2.26 and FT4 ≤ 12.26 then Hypothyroid | 2/3 = 0.67 |
| 5 If TSH > 0.91 and FT4 ≤ 12.26 and FT4 ≤ 19.63 and FT3 ≤ 0.77 and FT3 ≤ 0.69 then Hypothyroid | 3/3 = 1.0 |

| Rules generated for Subclinical Hyperthyroid | Confidence |
|---------------------------------------------|------------|
| 6 If TSH ≤ 0.91 and FT4 > 12.26 and FT4 ≤ 19.63 and FT3 > 0.768 then Subclinical Hyperthyroid | 3/3 = 1.0 |

| Rules generated for Hyperthyroid | Confidence |
|---------------------------------|------------|
| 7 If TSH ≤ 0.91 and FT4 > 12.26 and FT4 > 19.63 then Hyperthyroid | 9/11 = 0.82 |
| 8 If TSH ≤ 0.91 and FT4 > 12.26 and FT4 > 19.63 and FT3 ≤ 0.768 and FT3 > 0.69 then Hyperthyroid | 3/3 = 1.0 |

| Rules generated for Euthyroid | Confidence |
|-------------------------------|------------|
| 9 If TSH > 0.91 and FT3 > 0.80 and TSH > 5.41 and FT3 > 1.34 then Euthyroid | 1/2 = 0.50 |
| 10 If TSH > 0.91 and FT3 > 0.80 and TSH ≤ 5.41 then Euthyroid | 17/19 = 0.89 |
| 11 If TSH > 0.91 and FT3 ≤ 0.80 and TSH ≤ 2.26 and FT3 ≤ 12.26 then Euthyroid | 2/3 = 0.67 |

![Figure 4: Decision Tree Generated by J48](image)

### 5 Conclusion

This study suggests the practice of decision tree method for the classification of the thyroid disease. To decide the finest classifier of thyroid disease, current decision tree algorithms are compared with respect to average classification accuracy, precision value, recall value, F-measure and mean absolute error. Experimental studies were employed with thyroid data collected from a hospital by a general surgeon. NBTree has the highest accuracy rate with the value of 75%. Also, Friedman Two Way Analysis of Variances by Ranks Test is performed in order to determine that there is any difference among the algorithms effect or not. With four algorithms and six performance metrics (ACC(%), MAE, PRE, REC, FME, and Kappa Statistic), it is seen that p value is below 0.05 (p=0.001<0.05) alpha level significance. It is concluded that the algorithms have differences. Hence, pairwise comparisons are done for eleven algorithms. 40% of
these significant comparisons include NBTree. The running times of the algorithms are compared. It is seen that J48 is more speedy and interpretable than the NBTree is. As a result, it is obtained that decision tree learning has an ability to make prediction in order to diagnose the thyroid disease of a patient. For the feature research, some intelligent techniques such as fuzzy logic, artificial neural networks or their hybrid techniques as adaptive neuro-fuzzy inference system can be used for classification of the thyroid disease.

Acknowledgements

Dr. Gozde Ulutagay is supported by TWAS (The Academy of Sciences for the Developing World)-COMSTECH (Committee on Scientific and Technological Cooperation) in part under Grant No.10-183-RG/ITC/AS_C.

References

[1] F. Temurtas, A comparative study on thyroid disease diagnosis using neural networks, Expert Systems with Applications, 36 (2009) 944-49. http://dx.doi.org/10.1016/j.eswa.2007.10.010

[2] G. Ulutagay, Modeling of Thyroid Disease: A Fuzzy Inference System Approach, Wulfenia Journal, 19 (1) (2012) 346-57.

[3] G. Zhang, L. V. Berardi, An investigation of neural networks in thyroid function diagnosis, Health Care Management Science, (1998) 29-37. http://dx.doi.org/10.1023/A:1019078131698

[4] R. Erol, S. Noyan Oğulata, Şahin C, et al; A Radial Basis Function Neural Network (RBFNN) Approach for Structural Classification of Thyroid Diseases, J Med Syst, 32 (2008) 215-20. http://dx.doi.org/10.1007/s10916-007-9125-5

[5] K. Hoshi, J. Kawakami, M. Kumagai, et al; An Analysis of Thyroid Function Diagnosis Using Bayesian-Type and SOM-Type Neural Networks, Chem. Pharm. Bull, 53 (12) (2005) 1570-74. http://dx.doi.org/10.1248/cpb.53.1570

[6] S. C. Werner, Classification of thyroid diseases. Report of the committee on nomenclature. American Thyroid Association, J Clin Endocrinol Metab, 29 (1969) 860-62. http://dx.doi.org/10.1210/jcem-29-6-860

[7] S. C. Werner, S. H. Ingbar, Diseases of the thyroid. In: Werner SC, Ingbar SH, eds. The thyroid: a fundamental and clinical text, fourth Ed (New York: Harper and Row 1978) 389-393.

[8] H. Studer, H. J. Peter, H. Gerber, Natural heterogeneity of thyroid cells: the basis for understanding thyroid function and nodular goiter growth, Endocr Rev, 10 (1989) 125-135. http://dx.doi.org/10.1210/edrv-10-2-125

[9] J. W. F. Elte, J. K. Bussemaker, K. A. Haa, The natural history of euthyroid multinodular goiter, Postgrad Med J, 66 (1990) 186-190. http://dx.doi.org/10.1136/pgmj.66.773.186
[10] P. K. Sharpe, H. E. Solberg, K. Rootwelt, et al; Artificial neural networks in diagnosis of thyroid function from vitro laboratory tests, Clin. Chem, 39 (1993) 2248-2253.

[11] G. Serpen, H. Jiang, L. Allred, Performance analysis of probabilistic potential function neural network classifier, In Proceedings of artificial neural networks in engineering conference, St. Louis, MO, 7 (1997) 471-76.

[12] M. Brameier, W. Banzhaf, A Comparison of Linear Genetic Programming and Neural Networks in Medical Data Mining, IEEE Transactions on Evolutionary Computation, 5 (1) (2001) 17.  
http://dx.doi.org/10.1109/4235.910462

[13] L. Ozyilmaz, T. Yildirim, Diagnosis of thyroid disease using artificial neural network methods, In: Proceedings of ICONIP’02 9th international conference on neural information processing, (2002) 2033-36.  
http://dx.doi.org/10.1109/ICONIP.2002.1199031

[14] L. Pasi, Similarity classifier applied to medical data sets, In International conference on soft computing, Helsinki, Finland & Gulf of Finland & Tallinn, Estonia, (2004).

[15] K. Hoshi, J. Kawakami, M. Kumagai, et al; An analysis of thyroid function diagnosis using Bayesian-type and SOM-type neural networks, Chemical & Pharmaceutical Bulletin; 53 (12) (2005) 1570-74.  
http://dx.doi.org/10.1248/cpb.53.1570

[16] K. Polat, S. Sahan, S. Gunes, A novel hybrid method based on artificial immune recognition system (AIRS) with fuzzy weighted preprocessing for thyroid disease diagnosis, Expert Systems with Applications, 32 (2007) 1141-47.  
http://dx.doi.org/10.1016/j.eswa.2006.02.007

[17] A. Keles, A. Keles, ESTDD: expert system for thyroid diseases diagnosis, Expert Syst. Appl, 34 (1) (2008) 242-46.  
http://dx.doi.org/10.1016/j.eswa.2006.09.028

[18] P. B. Khanale, S. P., Ambilwade, A fuzzy inference system for diagnosis of Hypothyroidism, Journal of Artificial Intelligence, 4 (1) (2011) 45-54.  
http://dx.doi.org/10.3923/jai.2011.45.54

[19] E. Dogantekin, A. Dogantekin, D. Avci, An expert system based on Generalized Discriminant Analysis and Wavelet Support Vector Machine for diagnosis of thyroid diseases, Expert Syst.Appl, 38 (1) (2011) 146-150.  
http://dx.doi.org/10.1016/j.eswa.2010.06.029

[20] H-L. Chen, B. Yang, G. Wang, et al; A three-stage expert system based on support vector machines for thyroid disease diagnosis, J Med Syst, 36 (2012) 1953-63.  
http://dx.doi.org/10.1007/s10916-011-9655-8

[21] P. Jaganathan, N. Rajkumar, An expert system for optimising thyroid disease diagnosis, Int. J. of Computational Science and Engineering, 7 (3) (2012) 232-38.  
http://dx.doi.org/10.1504/IJCSE.2012.048243
[22] L-N. Li, J-H. Ouyang, H-L. Chen, et al; A computer aided diagnosis system for thyroid disease using extreme learning machine, J Med Syst, 36 (5) (2012) 3327-37. http://dx.doi.org/10.1007/s10916-012-9825-3

[23] D-Y. Liu, H-L. Chen, B. Yang, et al; Design of an enhanced fuzzy k-nearest neighbor classifier based computer aided diagnostic system for thyroid disease, J Med Syst; 36 (5) 3243-54. http://dx.doi.org/10.1007/s10916-011-9815-x

[24] A. T. Azar, A. E. Hassanien, T-H. Kim, Expert System Based on Neural-Fuzzy Rules for Thyroid Diseases Diagnosis, Communications in Computer and Information Science, 353 (2012) 94-105. http://dx.doi.org/10.1007/978-3-642-35521-9_13

[25] V. Rawte, B. Roy, Thyroid disease diagnosis using ontology based expert system, International Journal of Engineering Research & Technology, 4 (6) (2015) 365-70.

[26] I. M. D. Maysanjaya, H. A. Nugroho, N. A. Setiawan, A comparison of classification methods on diagnosis of thyroid diseases, In Proceedings of International Seminar on Intelligent Technology and Its Applications (ISITIA), (2015) 89-92. http://dx.doi.org/10.1109/ISITIA.2015.7219959

[27] S. A. Biyouki, I. B. Turksen, M. H. F. Zarandi, Fuzzy rule-based expert system for diagnosis of thyroid disease, In: IEEE Conference on Computational Intelligence in Bioinformatics and computational Biology (CIBCB), Niagara Falls, Canada, (2015) 1-7. http://dx.doi.org/10.1109/CIBCB.2015.7300333

[28] V. Prasad, T. S. Rao, M. S. P. Babu, Thyroid disease diagnosis via hybrid architecture composing rough data sets theory and machine learning algorithms, Soft Computing, 20 (3) (2016) 1179-89. http://dx.doi.org/10.1007/s00500-014-1581-5

[29] D. Lavanya, K. Usha Rani, Performance Evaluation of Decision Tree Classifiers on Medical Datasets, International Journal of Computer Applications, 26 (2011) 0975-8887. http://dx.doi.org/10.5120/3095-4247

[30] J. Margret, B. Lakshmipathi, S. A. Kumar, Diagnosis of Thyroid Disorders using Decision Tree Splitting Rules, International Journal of Computer Applications, 44 (8) (2012) 0975-8887. http://dx.doi.org/10.5120/6287-8474

[31] D. T. Larose, Discovering knowledge in data: An introduction to data mining, John Wiley & Sons, (2005) 385. http://dx.doi.org/10.1002/0471687545

[32] J. R. Quinlan, Decision Trees as Probabilistic Classifiers, In: Proceedings of the Fourth International Workshop on Machine Learning, (1987) 31-37. http://dx.doi.org/10.1016/b978-0-934613-41-5.50007-6

[33] L. Breiman, J. Friedman, R. Olshen, et al; Classification and Regression trees, Chapman & Hall, London, (1984).
[34] J. R. Quinlan, C4.5: Programs for machine learning, San Mateo, California: Morgan Kaufmann, (1993).

[35] Y. Zhao, Y. Zhang, Comparison of decision tree methods for finding active objects, Advances of Space Research, 41 (2007) 1955-59.
  http://dx.doi.org/10.1016/j.asr.2007.07.020

[36] L. Rokach, O. Maimon, Classification Trees, Data Mining and Knowledge Discovery, Springer New York Dordrecht Heidelberg London, (2010) 149-174.

[37] R. Kohavi, Scaling up the accuracy of naïve bayes classifiers: a decision tree hybrid, In: Proc. of the Second International Conference on Knowledge Discovery and Data Mining, Portland, Oregon, (1996) 202-207.

[38] H. Shi, Best-first decision tree learning, (MSc Thesis, University Waikato, 2007, Hamilton NZ).

[39] G. Holmes, B. Pfahringer, R. Kirkby, et al; Multiclass Alternating Decision Trees, Proceedings of the 13th European Conference on Machine Learning (ECML), (2002) 161-172.
  http://dx.doi.org/10.1007/3-540-36755-1_14

[40] I. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, San Mateo, CA, (2000).

[41] W. Fan, H. Wangi, P. S. Yu, et al; Is a random model better? On its accuracy and efficiency, In: Proceedings of Third IEEE International Conference on Data Mining (ICDM), (2003) 51.
  http://dx.doi.org/10.1109/ICDM.2003.1250902

[42] L. Breiman, Random forests, Mach. Learning, 45 (1) (2001) 5-32.
  http://dx.doi.org/10.1023/A:1010933404324

[43] N. Landwehr, M. Hall, E. Frank, Logistic model trees, Mach. Learning, 59 (2005) 592-98.
  http://dx.doi.org/10.1007/s10994-005-0466-3

[44] J. Gama, Functional trees, Mach. Learning, 55 (3) (2004) 219-50.
  http://dx.doi.org/10.1023/B:MACH.0000027782.67192.13

[45] J. J. Oliver, D. Hand, Averaging Over Decision Stumps, in Machine Learning, In Proceedings of European Conference on Machine Learning (ECML), Catania, Italy, (1994).
  http://dx.doi.org/10.1023/10.1007/3-540-57868-4_61

[46] J. Han, M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann Publisher is an imprint of Elsevier, Waltham, USA, (2006).

[47] G. Ulutagay, F. Ecer, E. Nasibov, Performance Evaluation of Industrial Enterprises via Fuzzy Inference System Approach: A Case Study, Soft Computing, 19 (2015) 449-58.
  http://dx.doi.org/10.1007/s00500-014-1263-3