JBACHS
Journal of Basic and Clinical Health Sciences

J Basic Clin Health Sci 2024; 8: 336-348
https://doi.org/10.30621/jbachs.1284274
RESEARCH ARTICLE

DECISION TREE-BASED CLASSIFICATION APPROACH TO DISCOVER FACTORS AFFECTING VITAMIN D LEVEL WITH MACHINE LEARNING

Ceyda Unal[1], Cihan Cilgin[2], Suleyman Albas[3], Esra Meltem Koc[4]

[1] Dokuz Eylul University Faculty of Economics and Administrative Sciences, Department of Management Information Systems, Izmir, Türkiye

[2] Bolu Abant Izzet Baysal University, Faculty of Applied Sciences, Department of Management Information Systems, Bolu, Türkiye

[3] Family Health Center, Karabaglar No 17, Izmir, Türkiye

[4] Izmir Katip Celebi University, Faculty of Medicine, Department of Family Medicine, Izmir, Türkiye

ORCID: C.U. 0000-0002-5503-8124; C.C. 0000-0002-8983-118X; S.A. 0000-0002-6779-5309; E.M.K. 0000-0003-3620-1261

Corresponding author: Ceyda Unal, **E-mail:** ceyda.unal@deu.edu.tr
Received: 16.04.2023; **Accepted:** 25.04.2024; **Available Online Date:** 31.05.2024

Cite this article as: Unal C, Cilgin C, Albas S, Koc EM. Decision Tree-Based Classification Approach to Discover Factors Affecting Vitamin D Level with Machine Learning. J Basic Clin Health Sci 2024; 8: 336-348.

ABSTRACT

Purpose: Vitamin D level is emphasized as an important biomarker in determining risk factors for different diseases. Vitamin D is an important vitamin for human health and its deficiency is associated with serious health problems. Therefore, it is of great importance to detect vitamin D deficiency, which can be easily prevented and treated. The possible relationship between vitamin D deficiency and musculoskeletal pain, osteoporosis, diabetes mellitus, hypertension is frequently discussed in researches. Enhanced availability of health data and decreased data processing expenses facilitate the extraction of valuable patterns related to vitamin D from extensive datasets. To illustrate, decision trees are commonly used for explainability and explainable AI (XAI) purposes. In this research, it is aimed to analyze the factors in determining the vitamin D level and the decision rules related to it.

Materials and Methods: A descriptive framework based on one of the machine learning techniques, that is decision tree is followed. The data used to create the decision rules were obtained from volunteers between the ages of 18-85 who applied to Izmir Katip Çelebi University Atatürk Training and Research Hospital Infectious Diseases and Family Medicine Polyclinics and agreed to participate in the study between 01.03.2017 and 01.09.2017. The sample size was calculated as 172 with 80% power, 5% error margin using NCSS and PASS software. The following parameters were examined: AST, ALT, ALP, BUN, creatine, total protein, albumin 25 (OH) D, PTH, TSH, Ca, Mg, phosphate, uric acid and VDR gene polymorphism. An investigator-designed socio-demographic data questionnaire was administered in-person interviews with 172 participants as a consequence of the research conducted with that total number of individuals.The validity of the models were assessed according to “accuracy scores” for each model.

Results: It was observed that age, gender and laboratory test values are strong predictors for vitamin D level. As a result of two CART (Classification and Regression Trees) models, %90.47 and %95 predictive accuracy were observed respectively. In the first model, uric acid, age and creatine; in the second model TSH, ALP and smoking(yes) were the most important three biomarkers affecting vitamin D level.

Conclusion: The collected features give a comprehensive list of variables that influence vitamin D in the dataset under consideration. Important findings of the study include not only the identification of these variables, but also the effective categorization determination procedures. Final decision tree models were constructed using two distinct feature sets. The initial model was created with 12 features (Age, ALP, TSH, URICACID, PHOSPHATE, AST, Cigarette Consumption, CA, CREATIN, TOTALPROTEIN, MG, BUN) that had over 4% importance, resulting in a classification accuracy rate of 92.7%. The second model was built using all features in the dataset and achieved a classification accuracy rate of 88.37%. In contrast to previous research, the Age variable is the most influential factor within the scope of this dataset, which includes demographic information on patients and their existing disorders.

Keywords: Machine learning, decision trees, decision rules, vitamin D.

INTRODUCTION

Big data in healthcare is crucial for processing and analyzing vast volumes of data. Artificial intelligence advancements contribute to clinical decision support systems, revealing valuable patterns from health data. Electronic health records and clinical analytics enable the analysis of large-quantitative data for new insights. As a result of this, reducing cost in healthcare domain has been inevitable in United States and other countries (1). Machine learning methods are transforming healthcare by identifying biomarkers in laboratory tests, with Vitamin D being a crucial biomarker for various healthcare situations.

Vitamin D deficiency is currently on increasing incidence globally, with a systematic review finding a prevalence of serum 25(OH)D < 30 nmol/L at 15.7% from 2000 to 2022. Vitamin D insufficiency is prevalent in Turkey as a consequence of restricted sunshine exposure and dietary effects. In Turkey, Vitamin D deficiency are quite common due to limited sunlight exposure and dietary factors (2). The primary factor contributing to the widespread occurrence of vitamin D deficiency is the failure to recognize that sun exposure has been and remains the primary source of vitamin D for the majority of adults and children [33–35]. Very few substances contain vitamin D naturally. These consist of cod liver oil, mushrooms that have been sun-dried or subjected to sunlight, and oily fish such as herring, mackerel, and salmon [1, 25, 34]. Meat, such as beef and pork, may contain an amount of vitamin D in the form of 25(OH)D, which can be significant at times [36, 37]. An increasing quantity of muscle 25(OH)D3 is being produced by numerous poultry, pigs, and cows as a result of the incorporation of 25(OH)D3 into diverse animal feeds. In addition to consuming polar bear liver and oily salmon, civilizations residing in the far Northern and Southern latitudes also obtain vitamin D from seal and whale blubber and polar bear liver.(3) The possible relationship between vitamin D deficiency and musculoskeletal pain, osteoporosis, diabetes mellitus, hypertension, cardiovascular diseases, autoimmune diseases, sleep disorders, cancer and increased mortality is frequently discussed in researches. However, there are limited number of researches about the factors related to Vitamin D level from the machine learning perspective.

In the literature various studies focused on identifying risk factors and associations related to vitamin D deficiency in different populations. Several studies found that factors like black race, female gender, winter season, and hypoalbuminemia were strong predictors of vitamin D deficiency in dialysis patients using decision tree-based algorithms (4-5-6-7-8-9). It was also indicated that vitamin D deficiency in the cities of Mashhad and Sabzevar in the northeast of Iran using the decision tree method based on 14 characteristics. 70% of the participants that is 618 cases were used as a random training dataset to form the decision tree, while the remaining 30% that is 285 cases were used as a test dataset to evaluate the performance of decision making. Using the test dataset, sensitivity, specificity, accuracy and AUC values were obtained as 79.3%, 64%, 77.8% and 72%, respectively. A study consisting of 31540 data presented a framework based on rules in the Apriori algorithm. A total of 22 association rules were generated with an 80% confidence level using WEKA software. The rule with the highest confidence level (98%) highlighted that among 1199 female patients aged 18 to 35 with vitamin D deficiency. (4-5-6). In addition, logistic regression and the XGBoost algorithm also used for detecting factors related to vitamin D level. The XGBoost algorithm identified stroke severity, age, and 25-hydroxyvitamin D level as important predictors (ROC/AUC of 0.805 versus 0.746). (7). A study conducted in Birjand, Iran, analyzing a healthy population to identify risk factors for vitamin D3 deficiency among chronic hepatitis B (CHB) patients. The study included 292 CHB patients and 330 vaccinated individuals, with serum biochemical characteristics measured. Data mining techniques were used, with 60% of the data used for training using the decision tree method. The model's performance was evaluated using the Receiver Operating Characteristics (ROC) curve, which yielded a 78% accuracy rate. The prevalence of vitamin D3 deficiency was found to be 63% among CHB patients and 32.9% among healthy individuals. The study concluded that serum zinc levels are predictive variables for vitamin D3 deficiency and emphasized the high accuracy in predicting the risk of vitamin D3 deficiency (8). It was also stated that Vitamin D was highlighted as a factor that reduces the risk of COVID-19. Countries were categorized into low or high COVID-19 cases, deaths, or case fatality rates based on the 40th and 60th percentiles (9).

The framework presented is identical as mentioned in previous studies but more focused on rule-based approach. In this context, a descriptive study is conducted. One of the purposes of the research is

analyzing the factors in determining the vitamin D level and decision rules behind them. The second is benefiting from the advantage of decision trees in terms of explainability and explainable AI (XAI). In order to do this, the remaining of article will emphasize the materials and methods in detail and discuss the results. Suggestions for further studies at the end of the research, could shed light on various researches.

MATERIALS AND METHODS

Dataset

The data used within the scope of the research were composed of volunteers between the ages of 18-85 who applied to the Infectious Diseases and Family Medicine Outpatient Clinics of İzmir Katip Çelebi University Atatürk Training and Research Hospital between 01.03.2017 and 01.09.2017 and agreed to participate in the research. Ethical approval obtained by İzmir Kâtip Çelebi University, Non-interventional Clinical Research Ethics Committee (Decision Date: 18.11.2021, Number: 0470). As a result of the research conducted with 172 people in total, a socio-demographic data questionnaire prepared by the researchers using the face-to-face interview technique was carried out. Exclusion criteria for study included: autoimmune disease, metabolic bone disease, chronic kidney disease, chronic liver disease, thyroid-- parathyroid disease, diabetes mellitus, malignancy, alcoholism, immunosuppression, liver transplantation, pregnancy or breastfeeding, psychiatric disease that disrupts the ability to answer questions, using medication that vitamin D, calcium, hormone therapy, glucocorticosteroid, antituberculosis, antiepileptic the last six months. In addition, as seen in detail in Table 1, it has been obtained from various biochemical data. The sample size used in this study was determined based on model tests performed as each new observation was added. Data collection was discontinued after a certain model performance was partially (approximately 90%) achieved.

Important factor considered within the scope of the purpose of the research is to discover the factors that affect the vitamin D level, so the dataset used in the research has a wide variable set consisting of 46 independent and 1 dependent variable. During the data preparation process, 25 (OH) D continuous variable was determined as one class and those with lower than 10 ng / ml as determining the vitamin D level, which is primarily used as an independent variable, while those higher than this threshold value were determined as another class. According to Turkey Endocrinology and Metabolism Association (TEMD) Osteoporosis and Metabolic Bone Diseases Working Group, a 25(OH)D level of at least 20 ng/ml (50 nmol/L) is accepted as sufficient for maintaining bone health. For non-bone effects, a level ranging from 30 to 50 ng/ml (75 to 125 nmol/L) is called adequate. A level between 10 and 20 ng/ml (25 to 50 nmol/L) indicates vitamin D insufficiency, while a value below 10 ng/ml (25 nmol/L) indicates vitamin D deficiency (10) Thus, the value “10 ng/ml” was chosen as threshold. The dependent variable, which is a continuous variable, has been transformed into a binary variable for the purpose of the research. In addition, variables in categorical form are organized as binary variables.

As stated in Table 1, the ages of the participants vary between 19-85 with an average age of 41.5 ± 13.6 years, consisting of 80 men and 92 women. In line with the information obtained from the participants, while 0 was determined for those who do not smoke, the number of daily use of cigarettes was used for smokers. Individuals were assigned with alcohol use habits, while 0 was assigned for individuals who did not consume alcohol.

In addition, all variables used in the dataset for the participants' existing diseases were determined as binary variables and a value of 1 was assigned in the presence of disease, while the opposite was indicated as 0. According to the test results obtained in the biochemical data, it was included in the dataset as a continuous variable in the relevant variable range as shown in Table 1. In addition, the DVit variable, which is used as the target variable within the scope of this research, is a binary variable and 84 of the participants are below the specified threshold value and represented by 0 value, while 88 of them are above the threshold value and represented with a value of 1.

Feature Selection

In machine learning problems, the representation of data often uses many features and only a few of them represent the target variable (11). Feature selection reflects the process of discovering a subset of related features or attributes as dependent variables in a predictive model, thereby helping to reduce the overfitting of the model and increase the prediction accuracy (12). Especially, as it is within the scope of

Table 1. Information about the Features used in the scope of the Dataset

Feature	Type	Range
Age	Continuous	19-81
Gender	Binary	0,1
Cigarette Consumption	Continuous	0-40
Alcohol	Binary	0,1
HBV (Hepatitis B virus)	Binary	0,1
CAH (Congenital Adrenal Hyperplasia)	Binary	0,1
HT (Hypertension)	Binary	0,1
ASTHMA	Binary	0,1
GOR	Binary	0,1
GASTRITIS	Binary	0,1
PANICDIS	Binary	0,1
ANEMIA	Binary	0,1
MIGRAINE	Binary	0,1
HL (Hodgkin Lenfoma)	Binary	0,1
LDH (Lactate Dehydrogenase)	Binary	0,1
DEPRESSION	Binary	0,1
PPI (Proton Pump Inhibitor)	Binary	0,1
ANTIHT	Binary	0,1
INHALER	Binary	0,1
ANTIAGREAGAN	Binary	0,1
SSRI (Selective Serotonin Reuptake Inhibitor)	Binary	0,1
FE	Binary	0,1
NSAII	Binary	0,1
STATIN	Binary	0,1
TOTALPROTEIN	Continuous	6.3 – 8.7
ALBUMIN	Continuous	3.2 – 4.8
CREATINE	Continuous	0.5 -1.2
AST (Aspartat Aminotransferaz)	Continuous	10 – 60
BUN (Blood Urea Nitrogen)	Continuous	1 – 30
CA (Calcium)	Continuous	5.9 – 10.5
PHOSPHATE	Continuous	1.9 – 5.1
MG (Magnesium)	Continuous	1.6 – 3.7
TSH (Thyroid Stimulating Hormone)	Continuous	0.24 – 5.85
PTH (Parathormon)	Continuous	0.1 – 99
ALP (Alkalen Fosfataz)	Continuous	9.1 – 179
URICACID	Continuous	1 – 8.94
ALT (Alanine Aminotransferase)	Continuous	3 - 82
AA	Binary	0,1
Aa	Binary	0,1
Aa	Binary	0,1
TT	Binary	0,1
Tt	Binary	0,1
Tt	Binary	0,1
FF	Binary	0,1
Ff	Binary	0,1
Ff	Binary	0,1
BB	Binary	0,1
Bb	Binary	0,1
Bb	Binary	0,1
Dvit	Binary	0,1

this study, feature selection phase is critical both in increasing the classification performance in datasets with huge variable sets and in the discovery process of important variables in the existing dataset.

Although there are substantial amount of feature selection methods in the literature, these methods can differ according to the types of variables in the dataset, the target variable and the machine learning approach to be applied to the dataset. In this context, Classification and Regression Trees (CART) -- a type of decision tree -- was used in both the classification task of the dataset and the feature extraction. As pointed out in the literature (13,14), the use of decision trees in feature selection is common and positively affects the performance in classification or regression tasks. In line with the discovery of the important variables that affect the target variable, which is one of the main objectives of the research, the feature selection process followed a unique framework different from the decision tree approaches in the literature. Especially ignoring the time complexity, the focus has been on determining the variable dataset that affects the prediction performance. Particularly, in the decision tree and feature selection approaches in the literature, the feature set is considered as a whole and the decision tree is evaluated on the tree structure formed with the variables in this whole set of variables, while the effect of different variable set combinations on the prediction performance and therefore on the feature selection is ignored. As a result of this, evaluation was made with all possible combinations of variable sets determined with the approach used in the research. It would be more informative to list the approach used in feature selection in the following steps:

Primarily, variables are divided into specified clusters according to determined similarities. (For example, variables belonging to diseases in the dataset can be considered as a single set.)

With all possible combinations of these cluster variable groups, decision trees were created with k-fold cross validation with the CART method. (the k value was determined as 20)

The decision trees created were evaluated according to the accuracy metric and the results that provided a certain accuracy rate (87% specified) were selected, and the percentage of variables that were effective in the decision rules used in the formation of these trees were determined.

Finally, the final variable significance were calculated by taking the mean values of variable significance obtained from decision trees created as a result of each different variable set combination.

Thus, a variable selection decision was not made over a single set of variables, and the attitudes of variables that occur in all possible sets of variables were examined through a comprehensive examination. At this point, it should be emphasized again that this process takes a lot of time (about one month in our research) but offers a robust approach in terms of reliability. Consequently, the feature selection methodology suggested would also be a contribution to the literature, especially in the process of determining independent variables effective on the target variable rather than time complexity.

Considering the trees created as a result of k-fold cross validation in the specified feature selection process, approximately 72.577.600 trees were evaluated and feature importance given in Table 2 were obtained. As can be seen in Table 2, Age, ALP, TSH and Uric Acid values are more than 7% importance in determining the target variable (Vitamin D), respectively. All variables (features) not included in the table were not reported to the research framework, as their importance were lower than 1%.

Table 2. The Order of Feature Importance Used in the Context of Dataset

Feature	Importance
Age	0,07954
ALP	0,07923
TSH	0,07267
URICACID	0,07015
PHOSPHATE	0,06826
AST	0,06157
Cigarette Consumption	0,05716
CA	0,05538
CREATINE	0,05224
TOTALPROTEIN	0,04402
MG	0,04386
BUN	0,04285
ALBUMIN	0,03892
PTH	0,03756
Aa	0,03163
ALT	0,02777
Gender	0,02736
HBV	0,02217
Tt	0,02011

Decision Trees- CART (Classification and Regression Trees)

Decision trees are one of the frequently used methods in data mining and machine learning. Finance (15,16), education (17,18), real estate (19,20), energy (21,22) and many similar areas, as well as in healthcare (23–26) preferred as a data mining technique. Decision Trees represents a tree-nodes corresponding to the order of decision rules in the simplest terms (27).

Today, decision trees are popularly preferred by researchers in the field of data mining because it has the advantage of ease of interpretation and visualization (28), does not require a preliminary process with its non-parametric modeling structure (29), requires very little data preparation, can process both numerical and categorical data and perform very well with a large dataset in a short time (15). One of the advantages of decision tree analysis is that the relationship between the binary dependent variable and the related independent variables is clearly demonstrated using a tree structure (29). In other words, it can be considered as a white box structure. In particular, unlike Black Box-type algorithms such as neural networks; decision trees are a white-box type machine learning algorithm, which is highly beneficial in evaluating the results and discovering the occurrence patterns (decision-making logic) of the results (30). In this way, a complex decision-making process can be divided into a collection of simpler decisions and decision rules, that are generally simpler to interpret (31) and understandable can be created. Decision trees can basically be designed for two task processes: classification tree analysis and regression tree analysis (15,20,24). Decision trees developed with the recursive partioning process provide a high-power tool for the definition, classification, regression and prediction of data (19). Decision trees generate the classification or regression process by using a set of hierarchical rules on variables, organized in tree structure (32).

Decision tree is one of the various approaches that can be used to develop a classification model for multi-stage decision making (31). It creates a tree-like structure model using inductive reasoning, focusing on existing data records (24). For this purpose, the decision tree starts with a root node where users can act, and from this node, users divide each node recursively according to the decision tree learning algorithm (33). The attribute/variable/feature is first classified (branched) in terms of groups and then the next important one is reconsidered and classified under information gain (17). In decision tree algorithms, the dataset is divided into two or more subgroups that are mutually exclusive at each split. The goal is to produce subsets of data that are as homogeneous as possible with respect to the target (dependent) variable (29). While performing this division function, it is necessary to determine how to divide trees that separates the decision tree algorithms from each other. Today many different various splitting criteria such as Entropy, Twoing, Gini; Gini Index, which is a binary splitting criterion, is more frequently preferred in datasets with continuous variables and is also used in this research.

For use in classification and regression tasks, decision tree theory is well suited for making medical predictions and data analysis statements in the field. Although there are many decision tree algorithms such as, ID3, C4.5, C5, CART, Random Forest and CHAID (Chi-square Automatic Interaction Detection) in literature, each of these algorithms can be applied to different datasets for different purposes. In this paper, CART (34) method, which can work easily with continuous variables and can also be used in regression problems, was preferred. The structure of the CART algorithm takes the independent variables into account in terms of predictive power; therefore it serves as a powerful discovery tool to understand the basic structure of the data. This algorithm is basically a series of carefully prepared questions about the features of the data, and after an answer has been generated for a question, a subsequent question is asked until the class is determined on the observation. These questions can be framed in the form of a hierarchical structure of nodes and directed edges (35). The CART procedure performs "binary recursive partitioning". The term "binary partitioning" means that the master node is continuously divided into two child nodes, and the term "recursive" means that the process is repeated, treating each child node as a parent node in the next step. This process is repeated until further partitioning is impossible, that is, until leaf nodes are formed or limited by some criteria determined by the user (36).

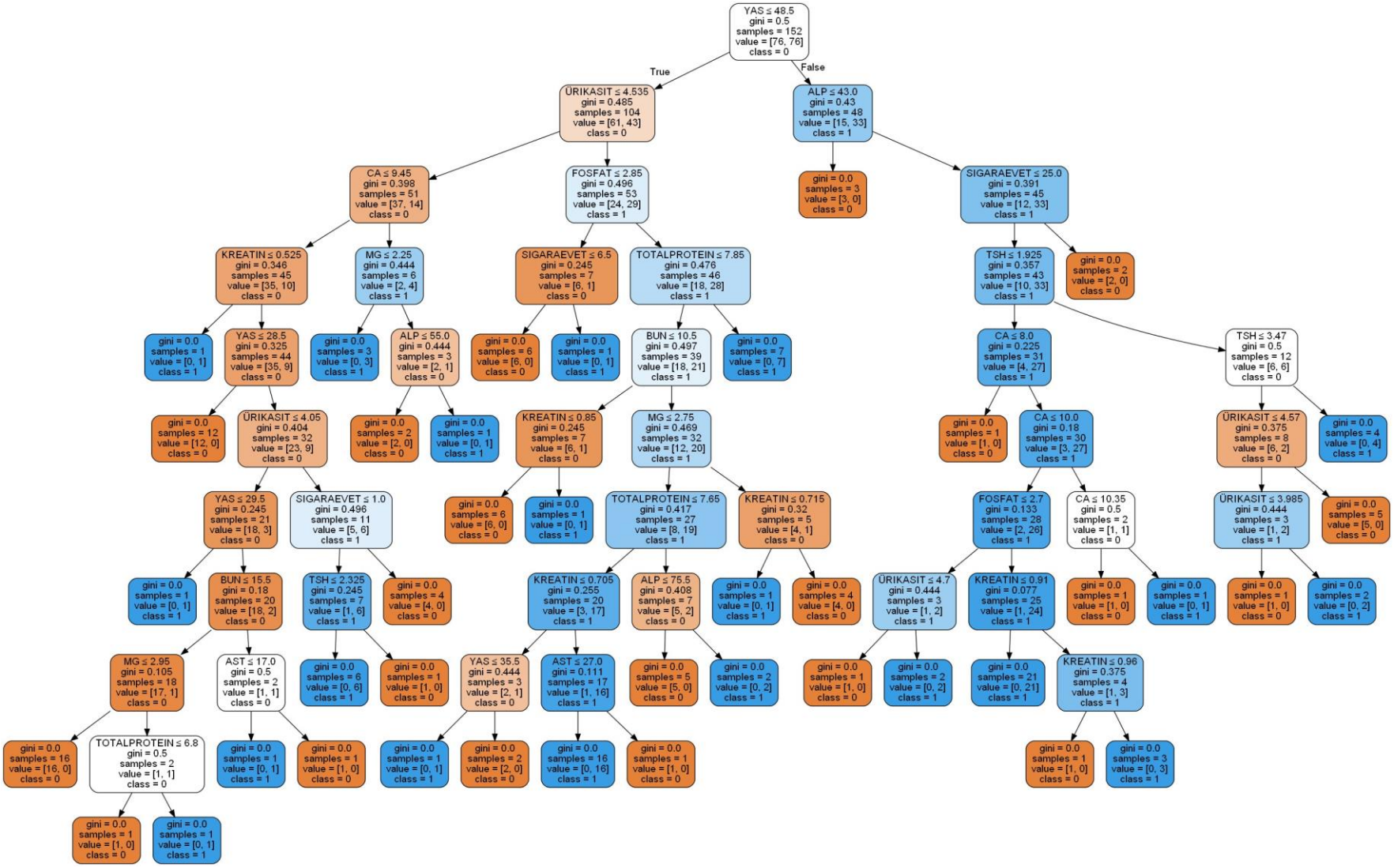

Figure 1. Initial CART Model (12 Features)

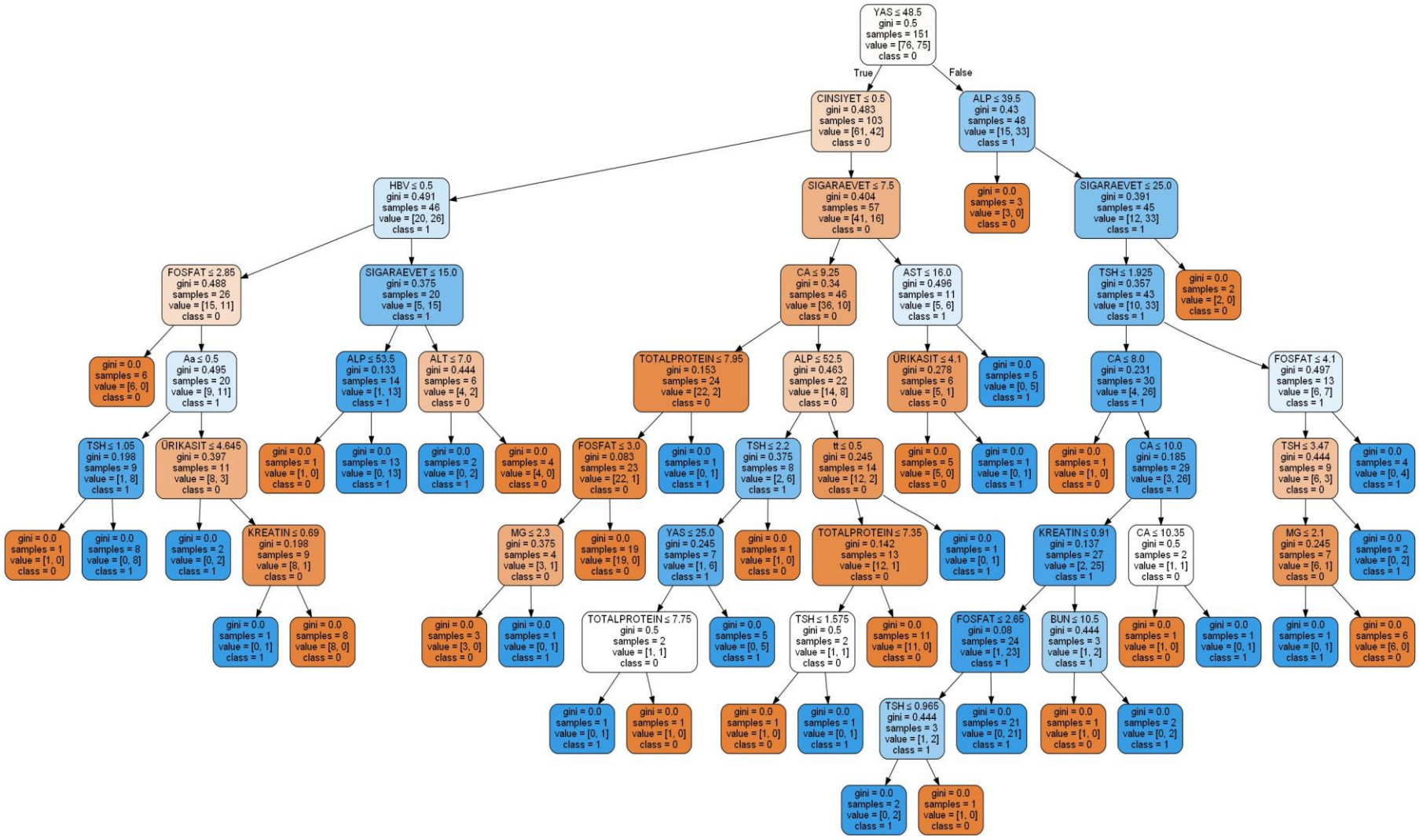

Figure 2. CART Model (All Features)

various splitting criteria such as Entropy, Twoing, Gini; Gini Index, which is a binary splitting criterion, is more frequently preferred in datasets with continuous variables and is also used in this research.

For use in classification and regression tasks, decision tree theory is well suited for making medical predictions and data analysis statements in the field. Although there are many decision tree algorithms such as, ID3, C4.5, C5, CART, Random Forest and CHAID (Chi-square Automatic Interaction Detection) in literature, each of these algorithms can be applied to different datasets for different purposes. In this paper, CART (34) method, which can work easily with continuous variables and can also be used in regression problems, was preferred. The structure of the CART algorithm takes the independent variables into account in terms of predictive power; therefore it serves as a powerful discovery tool to understand the basic structure of the data. This algorithm is basically a series of carefully prepared questions about the features of the data, and after an answer has been generated for a question, a subsequent question is asked until the class is determined on the observation. These questions can be framed in the form of a hierarchical structure of nodes and directed edges (35). The CART procedure performs "binary recursive partitioning". The term "binary partitioning" means that the master node is continuously divided into two child nodes, and the term "recursive" means that the process is repeated, treating each child node as a parent node in the next step. This process is repeated until further partitioning is impossible, that is, until leaf nodes are formed or limited by some criteria determined by the user (36).

RESULTS

Within the scope of the research, final decision tree models were carried out on two different feature set. Primarily, the first model was developed with 12 variables (Age, ALP, TSH, URICACID, PHOSPHATE, AST, Cigarette Consumption, CA, CREATIN, TOTALPROTEIN, MG, BUN) above 4% importance from the set of variables shown in Table 2. The second model was created with a total of 19 feature sets in the table. Within the framework of both models, 80% of the entire dataset was used as the training dataset, while 20% was used as the test dataset. Additionally, a 5-fold cross validation approach was adopted in the study to evaluate the model training results. In addition, the maximum depth of the decision tree is limited to 10 to prevent the complexity of the rules created by decision trees. Thus, the tree was completed after 10 branches. The tree was splitted according to Gini index as previously stated. In consequence of model implemented with the feature set within the first model, a high classification accuracy rate of 92,7% was achieved. The decision tree structure obtained as a result of this model is shown in Figure 1. As shown in Table 2, the “Age” feature has been assigned as the root node within the scope of the model.

Within the second model, a relatively high classification accuracy rate of 88.37% was achieved as a result of the model performed with the set of features included in the model. Although this accuracy rate is lower compared to the first model, it proves that the increase in the number of feature sets affects the model prediction performance as mentioned in the feature selection process. The decision tree structure obtained as a result of this model is shown in Figure 2. As in the first model, the “Age” feature, which is determined as the most important feature, was formed as the root node.

Alternative machine learning approaches were also applied within the scope of this study. In this way, it is possible to compare the results obtained with alternative methods, and the robustness of the results obtained can also be tested. For this purpose, analyzes were carried out with XGBooost, Random Forest (RF) and Support Support Vector Machine (SVM) models. The accuracy scores obtained as a result of these models are presented in Table 3.

For each model presented in Table 3, hyper parameter optimization was performed separately for two different feature sets with Grid-Search approach. As a result of this process, hyper-parameter values that provide the best model performance were selected. As seen in Table 3, SVM was the model with the lowest prediction performance. The main reason for this situation is the use of a tree-based approach in feature selection. Although RF and XGBoost showed relatively similar results, both models outperformed CART at a lower rate.

Although the results obtained within the scope of RF and XGBoost presented relatively better performance values, they did not provide a exceptionally increase in accuracy. Therefore, compared to CART, they require a large number of hyperparameter settings and therefore higher processing power and time. However, as another aim of this study, CART offers easy and fast use for many stakeholders. However, it should not be forgotten that the feature selection

Table 3. Accuracy scores obtained from alternative models

Models	Selected Hyper-Parameters	Accuracy Rate
SVM (1. Feature Set)	Kernel: Linear C: 1000 Gamma: 0.001	0,829
SVM (2. Feature Set)	Kernel: Linear C: 100 Gamma: 1	0,882
RF (1. Feature Set)	n_estimators: 50 Max_features: Sqrt Min_samples_leaf: 1 Max_depth: 2	0,931
RF (2. Feature Set)	n_estimators: 100 Max_features: Sqrt Min_samples_leaf: 2 Max_depth: 3	0,914
XGBoost (1. Feature Set)	Eta:0.01 Max_depth: 4 Subsample: 0,7	0,933
XGBoost (2. Feature Set)	Eta:0.01 Max_depth: 5 Subsample: 0,8	0,908

process within the scope of this study was carried out solely on the basis of the CART model. For this reason, it is expected that the accuracy rate provided by CART is high. Repeating similar feature selection processes for other tree-based approaches will further increase the accuracy rates of these models. In addition, modeling results performed with RF and XGBoost without any feature selection are much higher than both CART and SVM accuracy rates. This result is also an indicator of how effective the feature selection approach adopted in this study is, especially in improving the performance of CART.

DISCUSSION

This research aims to evaluate the factors influencing vitamin D levels and the corresponding the criteria for decision- Two separate feature sets were used to build the final decision tree models. The first model consisted of 12 features (Age, ALP, TSH, URICACID, PHOSPHATE, AST, Cigarette Consumption, CA, CREATIN, TOTALPROTEIN, MG, BUN) with overall importance above 4%, leading to a classification accuracy of 92.7%. The second model, utilizing all features in the dataset, had a classification accuracy rate of 88.37%. In the study, a significant relationship between high age, presence of chronic disease, being at university or higher education level, and high ALP and vitamin D levels was found. It is thought that this relationship may be due to the awareness of the retired and unemployed elderly population living in the province of Izmir, where the sociocultural level is high, about nutrition and benefiting from sunlight adequately. It is an expected situation that the average of vitamin D is determined to be higher in individuals with university and higher education level. Since the blood of the participants was collected within 2 months (March-April 2017), the seasonal variation was minimized. It was thought that vitamin D deficiency may have been detected more frequently, because the blood samples of the participants were taken after the winter season.

Moreover, unlike other studies in the literature, this study applies a high-precision feature selection process using decision trees. However, as presented in Table 2, the features obtained provide a complete list of variables that have an impact on vitamin D in the analyzed dataset. In addition to the identification of these variables, the decision rules that are effective in classification are also important findings of the study. Unlike the existing studies, the Age variable is the most important determinant within the scope of this dataset, which considers the demographic data of the patients as well as their existing diseases. Although Age has been used by many studies to determine vitamin D levels, it has not been identified as a factor, except for a few studies (8). Similar to other studies (9), Uric Acid (URICACID) and Calcium (CA) levels are other variables that affect the classification outcome. According to the empirical findings of the study conducted by (37) in which no

other variables were used except for various measurement values, ALBUMIN and ALT variables were the most influential variables on vitamin D levels, while these two variables were found to be relatively less influential variables in our study. In particular, this supports the conclusion that some demographic characteristics of the patients may be more effective as determinants of vitamin D levels. In addition to the highly interpretable findings of this study, the empirical results demonstrate classification performance with a high accuracy rate (95% accuracy).

The number of observations and variables used in this study constitute the main limitations. The data used in the study were conducted in a hospital environment, especially on people who performed certain laboratory tests. Therefore, the number of data is limited due to the data collected only from volunteers among those who performed these laboratory tests. Another limitation of this study is that different laboratory test results cannot be added as variables.

Studies on vitamin D prevalence and vitamin D cut-off value should be done in Turkey. In addition, this study needs to be repeated with new studies that measure vitamin D levels with a different method.

CONCLUSION

At the present time where data is highly apparent, healthcare services are also going through a big data revolution. Patients are also among the most critical elements of this ecosystem. Patients are constantly generating data and transferring their data to different applications. With regard to healthcare, it takes important steps towards personalized care, which is guided by an evidence-based approach to decision-making.

Artificial intelligence applications in healthcare bring about an important discussion. How these technologies can be included in the clinical workflow has become a critical issue studied by different researchers. As a result of this; two ways of positioning artificial intelligence in medicine / health are emerging: first, artificial intelligence is positioned as an aid for physicians and patients, second, and more radical, it replaces doctors as soon as it is sufficiently developed. The first is that artificial intelligence; as an irreplaceable component primarily in medicine, it assumes that human beings follow the principle of physicians, because it is above all a technology created by humans and humans are too complex structures to be analyzed from all aspects required by any artificial system. In this research, a framework supporting the first opinion for the purpose of determining Vitamin D level were proposed.

The research developed two decision tree models using two different feature sets. The first model had 12 variables with a 4% importance, while the second had 19 variables. 80% of the dataset was used as the training dataset, and 20% as the test dataset. A 5-fold cross validation approach was used to evaluate the model training results. The decision tree was split according to Gini index. The first model achieved a high classification accuracy rate of 92.7%, with the "Age" feature as the root node. The second model achieved a higher accuracy rate of 88.37%, indicating that increasing the number of feature sets affects model classification performance.

Based on these results, the decision tree method can serve as important and useful references in diagnosis for physicians to avoid the use of unnecessary medical supplies and improve healthcare quality. The empirical findings of this paper try to provide a reference index system for physicians in clinical diagnosis by using the decision tree, which is a machine learning approach. For example, it is possible that the factor rules generated from the decision tree model can be used in the judgment process to reduce human errors and avoid medical waste. In addition to the comprehensive examinations performed by specialist physicians, it is possible to provide a decision support to make the final diagnosis with higher accuracy with the information provided by this research, and this information can also be used to formalize and optimize the healthcare process.

Acknowledgements: The article was presented as an abstract at II. International Artificial Intelligence in Health Congress (2021).
Author Contributions: Concept- CC, CU; Design- CC; Supervision– EMK; Resource– EMK, SA; Materials– EMK, SA; Data Collection and/ or Processing- SA; Analysis and/or Interpretation- CC, CU; Literature Search- CC,CU EB; Writing– CC, CU, EMK; Critical Reviews- EMK.
Conflict of interest: None.
Ethical approval: Ethical approval obtained by Izmir Katip Celebi University, Non-interventional Clinical Research Ethics Committee (Decision Date: 18.11.2021, Number: 0470)
Funding: No financial funding.
Peer-Review: Externally peer-reviewed.

REFERENCES

1. Bates DW, Saria S, Ohno-Machado L, Shah A, Escobar G. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. Health Aff (Millwood) 2014;33(7):1123–1131.
2. Cui A, Zhang T, Xiao P, Fan Z, Wang H, Zhuang Y. Global and regional prevalence of vitamin D deficiency in population-based studies from 2000 to 2022: A pooled analysis of 7.9 million participants. Front Nutr 2023;10:1070808.
3. Holick MF. The vitamin D deficiency pandemic: Approaches for diagnosis, treatment and prevention. Rev Endocr Metab Disord 2017;18(2):153-165.
4. Bhan I, Burnett-Bowie Sam, Ye J, Tonelli M, Thadhani R. Clinical measures identify vitamin D deficiency in dialysis. Clin J Am Soc Nephrol CJASN 2010;5(3):460–467.
5. Gonoodi K, Tayefi M, Saberi-Karimian M, et al. An assessment of the risk factors for vitamin D deficiency using a decision tree model. Diabetes Metab Syndr Clin Res Rev 2019;13(3):1773–1777.
6. Kaya B, Günay A, Ozudogru, O. Analysis of the association between vitamin D deficiency and other diagnoses of patients by data mining techniques. Sakarya University Journal of Computer and Information Sciences 2020;3(1):51-59.
7. Kim C, Lee SH, Lim JS, et al. Impact of 25-hydroxyvitamin D on the prognosis of acute ischemic stroke: machine learning approach. Front Neurol 2020;11:37.
8. Osmani F, Ziaee M. Assessment of the risk factors for vitamin D3 deficiency in chronic hepatitis B patient using the decision tree learning algorithm in Birjand. Inform Med Unlocked 2021;23:100519.
9. Rahimi Z, Abdolvand N, Sepehri MM, Khavanin Zadeh M. The association of vitamin-D level with catheter-related-thrombosis in hemodialysis patients: A data mining model. J Vasc Access 2023;24(4):606-613
10. Turkish Journal of Endocrinology and Metabolism [Internet]. Osteoporosis and Metabolic Bone Diseases Diagnosis and Treatment Guide. [Accessed Date: 25 February 2024]. Available from www.temd.org.tr/files/OSTEOPOROZ_web.pdf.
11. Kira K, Rendell LA. A practical approach to feature selection. In: Sleeman D, Edwards P, editors. Machine Learning Proceedings. San Francisco (CA): Morgan Kaufmann; 1992. p. 249–256.
12. Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res 2003; 3: 1157-1182
13. Ratanamahatana C “ann”, Gunopulos D. Feature selection for the naive bayesian classifier using decision trees. Appl Artif Intell 2003;17(5–6):475–487.
14. Sugumaran V, Muralidharan V, Ramachandran KI. Feature selection using decision tree and classification through proximal support vector machine for fault diagnostics of roller bearing. Mech Syst Signal Process. 2007;21(2):930– 942.
15. Delen D, Kuzey C, Uyar A. Measuring firm performance using financial ratios: A decision tree approach. Expert Syst Appl. 2013;40(10):3970–3983.
16. Liu C, Hu Z, Li Y, Liu S. Forecasting copper prices by decision tree learning. Resour Policy. 2017 ;52:427–434.
17. Agarwal S, Pandey GN, Tiwari M. Data mining in education: data classification and decision tree approach. Int. J. e-Educ. e-Bus. e-Manag. e-Learn 2012; 2(2): 140.
18. Kolo KD, Adepoju SA, Alhassan J. A Decision Tree Approach for Predicting Students Academic Performance. Int. J. Edu. Mng Eng. 2015;5: 12-17.
19. Fan GZ, Ong SE, Koh HC. Determinants of house price: a decision tree approach. Urban Stud. 2006;43(12):2301–2315.
20. Shinde N, Gawande K. Survey on predicting property price. In: 2018 International Conference on Automation and Computational Engineering (ICACE). 2018 Oct 3-4. India p. 1–7.
21. Li X, Chan CW, Nguyen HH. Application of the neural decision tree approach for prediction of petroleum production. J Pet Sci Eng. 2013; 104:11–16.
22. Mikučionienė R, Martinaitis V, Keras E. Evaluation of energy efficiency measures sustainability by decision tree method. Energy Build. 2014;76:64–71.
23. Razavi AR, Gill H, Ahlfeldt H, Shahsavar N. Predicting metastasis in breast cancer: comparing a decision tree with domain experts. J Med Syst. 2007; 31(4):263–273.

24. Chang CL, Chen CH. Applying decision tree and neural network to increase quality of dermatologic diagnosis. Expert Syst Appl. 2009; 36(2):4035–4041.
25. Bayat S, Cuggia M, Rossille D, Kessler M, Frimat L. Comparison of bayesian network and decision tree methods for predicting access to the renal transplant waiting list. Stud Health Technol Inform. 2009;150:600–604.
26. Chaurasia V, Pal S, Tiwari BB. Chronic kidney disease: a predictive model using decision tree. Int. J. Res. Eng. Technol. 2019; 11(11): 1781-1794.
27. Singh D, Choudhary N, Samota J. Analysis of data mining classification with decision tree technique. Glob J. Comp Sci 2013; 13(13): 7-14.
28. Koh HC, Tan G. Data mining applications in healthcare. J Healthc Inf Manag JHIM. 2005;19(2):64–72.
29. Gandomi AH, Fridline MM, Roke DA. Decision tree approach for soil liquefaction assessment. Sci World J. 2013; 1-9.
30. Kurt AS, Cilgin C. Dış Ticaret Verileri İçin Kümeleme Analizi: Türkiye, Azerbaycan ve Kazakistan Örneği. Sosyoekonomi. 2021; 29(48): 511-540.
31. Safavian SR, Landgrebe D. A survey of decision tree classifier methodology. IEEE Trans Syst Man Cybern. 1991;21(3):660–674.
32. Aggarwal CC. Data mining: The textbook Springer International Publishing; 2015.
33. Singh S, Gupta P. Comparative study ID3, CART and C4.5 decision tree algorithm: A survey. Int. J. Adv. Sci; 27(27): 97-103.
34. Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and regression trees. Taylor & Francis; 1984. Chapman and Hall/CRC.
35. Sarkar S, Patel A, Madaan S, Maiti J. Prediction of occupational accidents using decision tree approach. In: 2016 IEEE Annual India Conference (INDICON). 2016 Dec 16-18. Bangalore, India.p. 1–6.
36. Waheed T, Bonnell RB, Prasher SO, Paulet E. Measuring performance in precision agriculture: CART—A decision tree approach. Agric Water Manag. 2006; 16;84(1):173–85.
37. Hoffmann G, Bietenbeck A, Lichtinghagen R, Klawonn F. Using machine learning techniques to generate laboratory diagnostic pathways—a case study. J Lab Precis Med. 2018; 3(6). 58.