Journal of Statistics & Management Systems
ISSN 0972-0510 (Print), ISSN 2169-0014 (Online)
Vol. 26 (2023), No. 2, pp. 317–330
DOI : 10.47974/JSMS-977

Risk factor analysis of COVID-19 with CART model

V. Deneshkumar *
R. Jebitha †
G. Jithu §
Department of Statistics
Manonmaniam Sundaranar University, Abishekapatti
Tirunelveli
Tamil Nadu 627012
India

Abstract

COVID-19 pandemic is presenting a great need for understanding the symptoms of people affected by the disease. The main objective of this work is implementation of CART methodology to determine the symptoms associated with COVID-19 disease. The decision tree demonstrates the classification route for detecting significant risk factors of COVID-19. The findings in this work will be helpful for medical decision makers and public health to prevent the spread of COVID-19.

Subject Classification: *62P10.*

Keywords: *Machine learning, CART model, Decision tree, Gini index and COVID-19.*

I. Introduction

Classification and Regression Trees (CART) assumes the independent samples to compute classification rules. It is used to explain a continuous or categorical dependent variable in terms of multiple independent variables. Decision tree are widely used in machine learning for classification and regression problems. Data is repeatedly partitioned into multiple sub-spaces by decision tree models in order to provide homogeneous outcomes

*Orcid Id: **0000-0001-8047-043X**

*E-mail: **vdenesh77@gmail.com (Corresponding Author)**
† E-mail: **jebitharajan1998@gmail.com**
§ E-mail: **jithujio123@gmail.com**

© Taru Publications

as possible in each final sub-space. Recursive partitioning is a technical term for this method [12]. CART analysis has become a useful tool in the analysis of public health data. Lemon, et al. [15] described the overview of CART analysis which estimated the percentage of the dependent variable among the strata and the logistic regression analysis. Marshall [19] has used the applications of classification trees in clinical epidemiology.

To forecast the risk of essential hypertension, Ture, et al. [28] evaluated the performance of classification techniques. Furthermore, they compared several statistical algorithms, decision trees and neural networks also. In a three-way analysis of smoker's data, Razi and Athappilly [23] compared nonlinear regression, NNs and CART models in order to assess prediction accuracy. Their results showed that NNs and CART models provided more accurate predictions than non-linear regression models. Nishida et al. [21] analyzed the periodontitis risk factors of personal risk with the poor lifestyle by using the CART method and multiple logistic regression analysis. For periodontitis, smoking and obesity are the independent risk indicators. Bel, et al. [2] presented some adaptations to the CART algorithm which is useful when the sampling pattern was irregular in the presence of clusters. They tested with two methods simultaneous and classical dataset and also highlighted their advantages and drawbacks. Chen et al. [6] have studied the data mining techniques to explore the risk factors of parenting stress in SPSS. The results of this work will be helpful to parents to prevent and reduce their stress. Chen, et al. [5] conducted a study based on nested case control design by using datamining techniques with Neural Network and decision tree to explore the risk factors of preterm. They provided the information for pregnant women to prevent the incidence of preterm. Mahmood et al. [18] have proposed methods for decision tree generation with accuracy. They also performed a comparative experiment with various real world data sets. The results showed the suitability and variability in knowledge discovery for moderate and large data sets.

Through the computerized analysis of children assessment data using Artificial Neural Networks (ANNs) and CART techniques, Yeh, et al. [30] have developed an intelligence model for early classification of real problems. De Nicola et al. [8] in classification and regression trees for getting the quality results they have used data envelopment analysis and also corrected the bias variables which is affecting health efficiency. Kuhn et al. [14] have discussed CART model for interaction between health-related variables in nursing research. Alizadehsani, et al. [1] have applied the several algorithms and measured the performance of each algorithm with data mining methods. Morgan [20] implemented the CART method in R and SAS package and discussed about bagging and boosting variation method. Gass, et al. [9] used the modified CART approach for the air pollution problem which is useful for understanding of complex multipollutant exposures. Loh,[17] reviewed classification algorithms and unbiased approaches with various useful applications. Hayes et al. [11] applied the CART and Random Forest methods for handling the missing data with Complete Case Analysis and Multiple Imputation. They suggested that bias and efficiency results from the cart analysis performed better than other methods.

Kong, et al. [13] have applied decision tree analysis for mortality prediction and they have evaluated the association between in-hospital mortality. Speiser, et al. [27] CART methodology gives the better performance in predicting in-hospital mortality. Zimmerman, et al. [32] estimated the probabilities for influenza with classification and regression trees model. And also performed the sensitivity analysis, multivariable regression analysis to identify the influenza among outpatients. Schilling, et al. [25] studied the patterns of primary prevention of cardiovascular disease with CART analysis. The receiver operating characteristics curves and optimized via pruning was used for performance evaluation. Several weight loss predictive models are discussed in Cheng et al. [7] study using incremental data with improved ROC curves. For radiation therapy patients with head and neck cancer, they examined the feasibility of CART prediction models for weight loss. Pergialiotis, et al. [22] investigated the diagnostic accuracy for predicting cancer in women with vaginal bleeding and the results were compared with Logistic Regression, ANN and CART Model. CART model finds the best optimum maximum depth of the tree based on the error analysis Ghiasi, et al. [10]. They have studied CART algorithm to evaluating the performance of graphical classifier method for the diagnosis of CAD. Zheng, et al. [31] used the Classification and Regression tree (CART) analysis to detect the potential risk factors associated to IFI, which is very useful for identifying the individuals who may be at high risk of dehydration in optimization of intervention programs. Rucci, et al. [24] discussed the CART and generalized linear model to stratify the diabetes patients into homogenous groups based on demographic and clinical characteristics. The prediction of COVID-19 transmission in real-time is crucial. Using time series data collected between January 22, 2020 and April 25, 2020 worldwide. Singh et al. [26] predicted the COVID-19 in terms of confirmed cases, deceased cases and recovered cases. Identification, deceased cases and recovery cases were explored using SVM. For COVID-19 data in India, Bhatnagar et al. [3] have performed descriptive statistical analyses. Different attributes are analyzed, such as age, gender, travel history and communication method. Patient transmission type and gender were significantly correlated. Tyasi, et al. [29] discussed the regression tree methods for prediction body weight and also predictive performances was measured by CART, CHAID and exhaustive CHAID algorithms. They found that CART was the best decision tree algorithm which gives the high predictive accuracy. As we known, very few studies used machine learning methods to explore factors of health problems. The purpose of this work is to implementation of CART methodology to determine the symptoms associated with Covid -19 disease.

II. Methodology

Using the hierarchical representation of the branches, a decision tree selects branching variables and methods according to the target setting. In order to explore or predict data, it is possible to prune decision tree models based on extracted classification rules. Furthermore, a hierarchical relationship can be discovered between the target variable and other variable. As part of attribute distribution, the decision tree predicts target variables and identifies root nodes to verify and trim data. A decision tree represents a test result in accordance with decision tree rules and methods. Leaf nodes display the distribution of target variables. It is possible to calculate the paths from a root node to a leaf node using a decision tree rule. By identifying a tree structure, decision trees can determine the attributes of target variable distributions that are highly influential. Classifying data can be accomplished through the selection of variables and the use of a hierarchical structure to identify the target and prediction model.

CART Algorithm:

Breiman et al. [4] proposed the CART method, which uses a binary segmentation approach. To distinguish the splitting condition, the Gini index is used. The data is split into two parts by each splitting method. Subset is then further divided to find the test attribute and also the splitting procedure is repeated when there is no further splitting possible. Pruning occurs after the CART algorithm has been learned with the overall error rate. The most efficient classification gives the smallest tree. The CART algorithm can be applied to a target variable that represents both continuous and categorical data. A regression tree used for the continuous target variable and classification tree used for the categorical target variable. For each node, the CART algorithm chooses a dynamic threshold value. A binary decision tree is created by dividing the data at each node using a single input variable function. In order to estimate the risk factors, the Gini index is used. A significant number of risk factors indicates the presence of numerous categories in the data. Assume that there are n samples in the data set G. Pj represents the relative probability from j show up in G.

$$Gini(G) = 1 - \sum_{j=1}^{n} pj^2 \qquad ...(1)$$

Suppose that data set G split as two subsets G1 and G2 on A, where n1 and n2 denotes the number of data in G1 and G2. Also, n = n1 + n2.

$$Gini(G) = \frac{n_1}{n} Gini(G_1) + \frac{n_2}{n} Gini(G_2) \qquad ...(2)$$

Gini index's are designed to determine which categories in the data have the greatest number of instances compared to other categories at different nodes. Low Gini values indicate an inconsistent category distribution in the sample. It reveals that in subset formed by the splitting point, category purity and capacity to compare among different categories both increases (Lin and Fan [16]).

III. Results and Discussions

The data classifies the patients on the bases of whether or not the patient has a symptom associated to COVID-19 disease. The data is obtained from a report published by WHO in June 2020. The collected data comprises the information of 21 risk factors associated to COVID-19 disease. This data consists of 5434 patients and classifies whether or not the patient has the symptom and classifies the person with or without COVID-19 by yes or no. There are 21 independent parameters. The independent parameters included were fever, contact with Covid patient, breathing problem, sore throat, dry cough, running nose, heart disease, lung disease, headache, diabetes, hyper tension, gastrointestinal, asthma, fatigue, wearing masks, visited public exposed places, abroad travel, attended large gathering, exposure to public places and others. Categorical binary variables were converted to numeric before entering into the model.

Dataset Splitting

For the purpose of developing CART, the COVID-19 dataset is divided training and testing data. The model is trained on 80% of the data, while it is tested on 20%. The structure of the CART algorithm depends upon the errors of the training samples.

Logistic Regression Model for Diagnosis of Covid -19

In a situation in which the dependent variable is dichotomous, binary logistic regression is appropriate. Our data contains COVID-19 as the dependent variable and COVID-19 symptoms as the independent variables the binary logistic regression model is fitted and the results are summarized in the Table 1.

The estimates of the model coefficients, their standard errors (SEs), Wald statistic, odds ratio and its associated significant p-values are presented in Table 1. The significant p-value computed with respect to the Wald test statistic for each predictor in presented in Table. It is obvious that there is a significant relationship between the variables dry cough, fever, running nose, breathing problem, sore throat, hypertension, asthma

Table 1

Coefficients of the logistic regression model

	B	S.E	Wald	df	p-value	Odds Ratio	95% C.I for Odds ratio	
							Lower	Upper
Breathing Problem	2.386	0.113	443.556	1	0.000	10.865	8.702	13.566
Fever	1.717	0.121	200.713	1	0.000	5.565	4.389	7.057
Dry Cough	2.750	0.122	504.445	1	0.000	15.650	12.310	19.896
Sore throat	1.966	0.109	323.280	1	0.000	7.140	5.763	8.846
Running Nose	-0.218	0.105	4.311	1	0.308	0.804	0.654	0.988
Asthma	-0.002	0.105	0.000	1	0.005	0.998	0..812	1.226
Lung Disease	0.074	0.105	0.499	1	0.023	1.077	0.876	1.324
Headache	0.011	0.106	0.012	1	0.914	1.012	0.822	1.245
Heart Disease	0.092	0.105	0.778	1	0.378	1.097	0.893	1.346
Diabetes	0.137	0.104	1.723	1	0.010	1.147	0.935	1.407
Hyper Tension	0.539	0.108	25.155	1	0.000	1.715	1.389	2.118
Constant	-4.380	0.217	408.362	1	0.000	0.013		

and lung disease. However, variables heart disease, running nose and headache not as statistically significant as the values are larger than 0.05. Therefore, it can be inferred as the overall model is significant. While all the other symptoms are statistically significant at 95% confidence interval. Hence, the sample information provides larger evidence for rejecting H0. Therefore, it can be inferred as the fitted model has overall significant. The value of the Hosmer-Lemeshow test statistic is 353.673 with significant p-value of 0.519. Therefore, the study model was quite a good fit. The classification table explains 92.3% of the patients who had the symptoms where correctly predicted to have the disease of the patients with specificity of 98.0% and sensitivity of 68.4%.

Classification and Regression Tree for Diagnosis of COVID-19

A classification and regression tree (CART) is a predictive model that demonstrates how the values of an outcome variable can be predicted based on the values of other variables. The CART output includes an output predictor variable for every fork and an outcome prediction node for each end node. In this case, the outcome variable is whether the subject has COVID-19 or not and the predictor variables are various symptoms of COVID-19.

First a tree is constructed using all the variables without any control over the depth and the number of nodes the tree obtained is a very complex tree and is very hard to interpret. Then the above data split into two parts as a training set and a testing set and it is split in such a way that the testing part contains 70% of the data and training part contains 30% of the

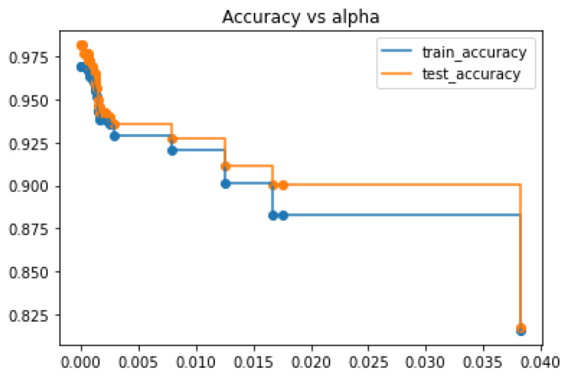

Figure 1

Accuracy for training and testing data

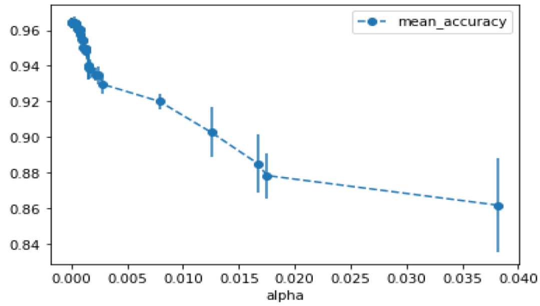

Figure 2

Mean accuracy

data. Cost complexity pruning can simplify the whole process of finding a smaller tree that improves the accuracy and also reduce the complexity of the tree. First, we construct the accuracy vs alpha graph to check the accuracy values of testing and training data.

From the Figure 1 shows that the accuracy of testing and training data set is almost similar throughout the graph. In Figure 2, the alpha values are between 0.001 and 0.002, the tree has very high accuracy value around 97.43% and the corresponding alpha value is 0.001002 which is perfectly pruned tree with a very high accuracy and less complexity.

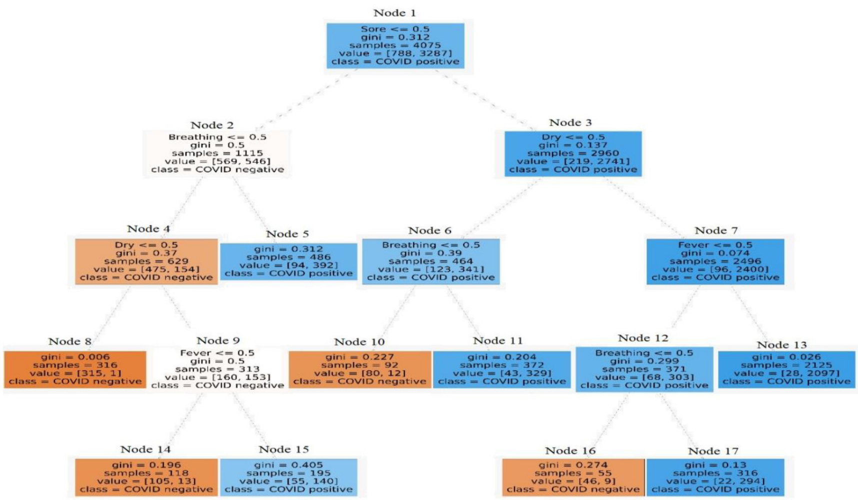

Figure 3

Classification and regression tree for risk factors of Covid- 19

The final pruned tree only took four predictor variables (sore throat, breathing problem, fever, dry cough) in constructing the tree indicating these are the most commonly associated symptoms of a COVID-19 positive patients. In Figure 3, the blue colored nodes indicate that the portionis corona positive and the orange shades indicate that the portionis COVID-19 negative.

In the primary split, the predictor variable Sore throat is used for the primary split and total of 4075 samples were considered and 3287 of them where positive and the split has a Gini index of 0.312. Secondary split consists of 2 nodes both of them are internal nodes leading to other variables node 2 has breathing as the variable with Gini index value of around 0.5 a total of 1115 samples were considered of which 569 were classified as negative and 546 where classified as positive. The node 3 has dry cough as the predictor variable with a Gini index value of 0.137, a total of 2960 samples were considered 2741 of them where positively identified. The third split gives rise to four nodes three of them are internal node and on leaf node.

Node 4 and node 5 arises as a result of splitting node number 2. Node 4 has again dry cough as the predictor variable with a gini index value of 0.37 and a total of 629 samples were considered where 154 where only identified as positives. Node 5 arises as the leaf node of node 2 with a gini index value of 0.5. Node 6 and node 7 arises as a internal node of node number 3. Node 6 has breathing problem as a predictor variable, a gini index of 0.39, a sample of 464 was considered among which 341 were identified as positives. Node 7 has fever as the predictor variable a sample of 2496 where considered among which 2400 were positively identified and the node has gini index of 0.074. The fourth split has 6 nodes among which only 2 are internal nodes and the rest of them are terminal nodes. Node 4 gives rise to node 8 and node 9, node 6 gives rise to node 10 and node11, node 7 gives rise to node 12 and node 13. Node 8 is a terminal node of node 6 with a gini index of 0.006. Node 9 has fever as the predictor variable with a gini index of 0.5, a total of 313 samples were considered and 153 of them were identified as positives. Node 10 is again a terminal node of node 6 with a gini index of 0.227. Node 11 is again a terminal node of node 6 with a gini value of 0.204. Node 12 is internal node of node 7 it has breathing as the predictor variable and a gini index of 0.299, total of 371 samples were considered in the node of which 303 were determined positive. Node 13 is a terminal node of node 7 with a gini index of 0.026. Finally, the fifth split has 4 nodes 4 of them are terminal nodes. Nodes 14 and 15 are terminal nodes of node 9 with a gini

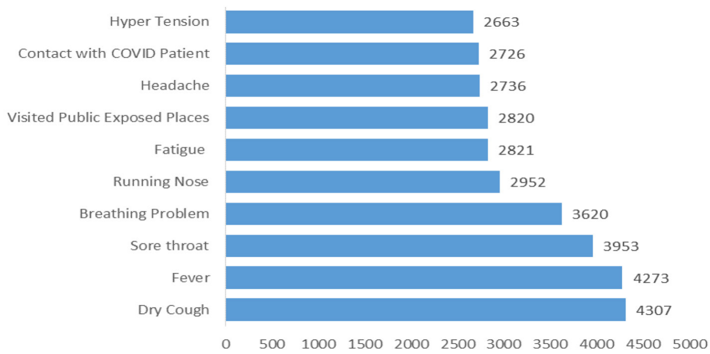

Figure 4

Feature importance

index of 0.196 and 0.405 respectively. Nodes 16 and 17 are terminal nodes of node 12 with gini index values of 0.274 and 0.13 respectively.

The feature importance Figure 4 shows which symptom has the high risk for affecting COVID-19 disease and which symptom has the low risk of getting affected by COVID-19 disease. From the future importance breathing problem, heart disease, fever, dry cough, abroad travel, contact with Covid patient, sore throat, large gathering, hyper tension, working in public exposed places have high risk of getting affected by COVID-19 disease.

IV. Conclusion

In this work, the CART methodology was used for identifying and estimating the risk factors of Covid -19 disease. Additionally, it can be used for monitoring and evaluating COVID-19 disorders, as well as for preventing and controlling them. Decision tree shows the classification route for detecting the significant risk factors of Covid -19 disease. The high accuracy was obtained from the CART Model with a classification rate of 97% when it was evaluated against the test set. The findings in this study are helpful for medical decision makers and social workers to prevent and reduce their risk of affecting corona virus.

Acknowledgements

The second author thanks the Manonmaniam Sundaranar University, Abishekapatti, Tirunelveli, Tamil Nadu for providing financial assistance.

under the scheme of Bharat Rathna Dr. M.G. Ramachandran Centenary Research Fellowship to carry out this work.

References

[1] Alizadehsani, R., Habibi, J., Hosseini, M. J., Mashayekhi, H., Boghrati, R., Ghandeharioun, A., ... & Sani, Z. A. (2013). A data mining approach for diagnosis of coronary artery disease. *Computer methods and programs in biomedicine,* 111(1), 52-61.

[2] Bel, L., Allard, D., Laurent, J. M., Cheddadi, R., & Bar-Hen, A. (2009). CART algorithm for spatial data: Application to environmental and ecological data. *Computational Statistics & Data Analysis,* 53(8), 3082-3093.

[3] Bhatnagar, V., Poonia, R. C., Nagar, P., Kumar, S., Singh, V., Raja, L., & Dass, P. (2021). Descriptive analysis of COVID-19 patients in the context of India. *Journal of Interdisciplinary Mathematics*, *24*(3), 489-504.

[4] Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). Classification and regression trees. CRC press.(Book).

[5] Chen, H. Y., Chuang, C. H., Yang, Y. J., & Wu, T. P. (2011). Exploring the risk factors of preterm birth using data mining. *Expert systems with applications,* 38(5), 5384-5387.

[6] Chen, H. Y., Hou, T. W., Chuang, C. H., & TBPS Research Group. (2010). Applying data mining to explore the risk factors of parenting stress. *Expert Systems with Applications*, 37(1), 598-601.

[7] Cheng, Z., Nakatsugawa, M., Hu, C., Robertson, S. P., Hui, X., Moore, J. A., ... & Quon, H. (2018). Evaluation of classification and regression tree (CART) model in weight loss prediction following head and neck cancer radiation therapy. *Advances in Radiation Oncology*, *3*(3), 346-355.

[8] De Nicola, A., Gitto, S., & Mancuso, P. (2012). Uncover the predictive structure of healthcare efficiency applying a bootstrapped data envelopment analysis. *Expert Systems with Applications*, 39(12), 10495-10499.

[9] Gass, K., Klein, M., Chang, H. H., Flanders, W. D., & Strickland, M. J. (2014). Classification and regression trees for epidemiologic research: an air pollution example. *Environmental Health*, 13(1), 1-10.

[10] Ghiasi, M. M., Zendehboudi, S., & Mohsenipour, A. A. (2020). Decision tree-based diagnosis of coronary artery disease: CART model. Computer methods and programs in biomedicine, 192, 105400.

[11] Hayes, T., Usami, S., Jacobucci, R., & McArdle, J. J. (2015). Using Classification and Regression Trees (CART) and random forests to analyze attrition: Results from two simulations. *Psychology and aging*, *30*(4), 911.

[12] Kassambara, A. (2018). *Machine learning essentials: Practical guide in R*. Sthda.

[13] Kong, A. L., Pezzin, L. E., & Nattinger, A. B. (2015). Identifying patterns of breast cancer care provided at high-volume hospitals: a classification and regression tree analysis. *Breast cancer research and treatment*, 153(3), 689-698.

[14] Kuhn, L., Page, K., Ward, J., & Worrall-Carter, L. (2014). The process and utility of classification and regression tree methodology in nursing research. *Journal of Advanced Nursing*, *70*(6), 1276-1286.

[15] Lemon, S. C., Roy, J., Clark, M. A., Friedmann, P. D., & Rakowski, W. (2003). Classification and regression tree analysis in public health: methodological review and comparison with logistic regression. *Annals of Behavioral Medicine*, *26*(3), 172-181.

[16] Lin, C. L., & Fan, C. L. (2019). Evaluation of CART, CHAID and QUEST algorithms: a case study of construction defects in Taiwan. *Journal of Asian Architecture and Building Engineering*, *18*(6), 539-553.

[17] Loh, W. Y. (2014). Fifty years of classification and regression trees. *International Statistical Review*, 82(3), 329-348.

[18] Mahmood, A. M., Imran, M., Satuluri, N., Kuppa, M. R., & Rajesh, V. (2011, December). An improved CART decision tree for datasets with irrelevant feature. In International Conference on Swarm, Evolutionary and Memetic Computing (pp. 539-549). Springer, Berlin, Heidelberg.

[19] Marshall, R. J. (2001). The use of classification and regression trees in clinical epidemiology. *Journal of Clinical Epidemiology*, *54*(6), 603-609.

[20] Morgan, J. (2014). Classification and regression tree analysis. Technical Report No. 1. Boston University of Public Health, 298 (https://www.bu.edu/sph/files/2014/ 05/MorganCART.pdf).

[21] Nishida, N., Tanaka, M., Hayashi, N., Nagata, H., Takeshita, T., Nakayama, K., ... & Shizukuishi, S. (2005). Determination of smoking

and obesity as periodontitis risks using the classification and regression tree method. *Journal of Periodontology*, *76*(6), 923-928.

[22] Pergialiotis, V., Pouliakis, A., Parthenis, C., Damaskou, V., Chrelias, C., Papantoniou, N., & Panayiotides, I. (2018). The utility of artificial neural networks and classification and regression trees for the prediction of endometrial cancer in postmenopausal women. *Public Health,* 164, 1-6.

[23] Razi, M. A., & Athappilly, K. (2005). A comparative predictive analysis of neural networks (NNs), nonlinear regression and classification and regression tree (CART) models. *Expert systems with applications,* 29(1), 65-74.

[24] Rucci, P., Avaldi, V. M., Travaglini, C., Ugolini, C., Berti, E., Moro, M. L., & Fantini, M. P. (2020). Medical Costs of Patients with Type 2 Diabetes in a Single Payer System: A Classification and Regression Tree Analysis. PharmacoEconomics-open, 4(1), 181-190.

[25] Schilling, C., Mortimer, D., Dalziel, K., Heeley, E., Chalmers, J., & Clarke, P. (2016). Using Classification and Regression Trees (CART) to identify prescribing thresholds for cardiovascular disease. *Pharmacoeconomics,* 34(2), 195-205.

[26] Singh, V., Poonia, R. C., Kumar, S., Dass, P., Agarwal, P., Bhatnagar, V., & Raja, L. (2020). Prediction of COVID-19 corona virus pandemic based on time series data using Support Vector Machine. *Journal of Discrete Mathematical Sciences and Cryptography*, *23*(8), 1583-1597.

[27] Speiser, J. L., Karvellas, C. J., Shumilak, G., Sligl, W. I., Mirzanejad, Y., Gurka, D., ... & Kumar, A. (2018). Predicting in-hospital mortality in pneumonia-associated septic shock patients using a classification and regression tree: a nested cohort study. *Journal of intensive care*, 6(1), 1-10.

[28] Ture, M., Kurt, I., Kurum, A. T., & Ozdamar, K. (2005). Comparing classification techniques for predicting essential hypertension. *Expert Systems with Applications*, 29(3), 583-588.

[29] Tyasi, T. L., Eyduran, E., & Celik, S. (2021). Comparison of tree-based regression tree methods for predicting live body weight from morphological traits in Hy-line silver brown commercial layer and indigenous Potchefstroom Koekoek breeds raised in South Africa. *Tropical Animal Health and Production*, 53(1), 1-8.

[30] Yeh, Y. L., Hou, T. H., & Chang, W. Y. (2012). An intelligent model for the classification of children's occupational therapy problems. *Expert Systems with Applications*, 39(5), 5233-5242.

[31] Zheng, H., Fei, J., Zhang, L., Zhou, W., Ding, Z., & Hu, W. (2020). Risk factor analysis of insufficient fluid intake among urban adults in Wuxi, China: a classification and regression tree analysis. *BMC public health*, 20(1), 1-10.

[32] Zimmerman, R. K., Balasubramani, G. K., Nowalk, M. P., Eng, H., Urbanski, L., Jackson, M. L., ... & Wisniewski, S. R. (2016). Classification and Regression Tree (CART) analysis to predict influenza in primary care patients. *BMC infectious diseases*, 16(1), 1-11.

Received November, 2021

Revised June, 2022