COMPUTATIONAL INTELLIGENCE AND MACHINE LEARNING
e-ISSN: 2582-7464

Computational Intelligence and Machine Learning
Vol-3, Issue-1, April 2022

Predictive Analytics of Selected Datasets using DTREG Data Mining Tool

Megha N[1*], Sandhya KV[2], Jithu Jerin James[3], Aksa Alex[4]

[1, 2, 3, 4] Faculty of Pharmacy, M S Ramaiah University of Applied Sciences, Bengaluru, Karnataka, India.
*Corresponding Author Email: [1] meghagowdadec90@gmail.com

Abstract

Predictive analytics is making a significant wave in healthcare Industry. Predictive analytics is an analytics offshoot which helps to make future predictions, resulting in more informed decisions. Data is central to accurate predictions. Several concepts like Data Mining, AI (Artificial Intelligence), Machine Learning and statistics need to work in tandem to ensure precise predictions. The main aim of the research work was to analyse the datasets of selected diseases using the DTREG data mining tool. Two datasets namely Alzheimer's and Breast Cancer were taken from a public repository and analysed. Various algorithms namely single tree, decision tree, tree boost, support vector machine and neural network were studied. The results obtained were interpreted to understand which algorithm works best in each case. Also, the important predictors in each study were recorded. Interpretation of Alzheimer's and breast cancer data using DTREG revealed neural network as the best algorithm. The significant predictors for Alzheimer's were estimated as total intracranial blood volume, clinical dementia rating and age, and for breast cancer were uniformity of cell size, cell shape, benign and malignant and clump thickness. Data mining, artificial Intelligence and machine learning can thus be of very good help in determining the line of treatment to be followed by extracting knowledge from such suitable databases.

Keywords

Algorithms, Alzheimer's, Breast Cancer, DTREG.

INTRODUCTION

Predictive analytics is a branch in the domain of advanced analytics. It is utilized in prognosticating the future events. It analyzes the current and historical data to make predictions about the future by employing the techniques from statistics, data mining, machine learning, and artificial intelligence. Predictive analysis can create more efficient and effective health systems. There are many applications of predictive analysis which includes, creating a prognosis score using patient records, genetic screening which can help to determine possible diseases, faster and more accurate interpretation of medical images like X-rays. Since a lot of patient data is available, even from wearables, predictive analysis can also help with accessing crucial patient data of remote patients and predict their hospital visits, admissions, and emergencies (proactive care). Predictive analytics can handle large data sets, for example, cohort data. With predictive analysis, it is possible to establish the general health of a community [1]-[3].

Data mining is a process of computing models or design in large collection of data using the steps - exploration, pattern identification and deployment. It is concerned together with the method of computationally extracting unknown knowledge from vast sets of data. Data mining can be used to extract knowledge by analysing and predicting some diseases. Data mining applications in health care can have a wonderful potential and effectiveness. It automates the process of finding predictive information in large databases. Disease prediction plays an important role in data mining. Health care data mining is an important task because it allows doctors to see which attributes are more important for diagnosis such as age, weight, symptoms, etc [3]-[8].

In this study, we selected DTREG, an open-source software and worked on data from selected databases. DTREG is a decision tree building software product that can be used for Predictive Modeling (Data Mining) and Forecasting. It can be used to predict values for future observations and also has full support for time series analysis. It accepts a dataset in the form of table containing number of rows, whose columns represent attributes/variables. One of the variables is the **''target variable''** whose value is to be modeled and predicted as a function of the **''predictor variables''.** The DTREG analyzes the data and generates a model showing how best it predicts the values of target variable based on the values of predictor variables. It builds classification and regression decision trees, neural networks (NN), support vector machine (SVM), gene expression programs, K-means clustering, discriminant analysis and logistic regression models that can describe data relationships. The significant features of DTREG includes ease of use, can build classification and regression trees, automatic tree pruning, surrogate splitter for missing data, visual display of the tree, acceptance of text data as well as numeric data, data transformation language (DTL) etc [5].

Alzheimer's and Breast cancer datasets were considered for the study. Alzheimer's disease is a progressive neurologic disorder that causes the brain to shrink (atrophy) and brain cells to die. Alzheimer's disease is the most common cause of dementia — a continuous decline in thinking, behavioural and social skills that affects a person's ability to function independently. The exact causes of Alzheimer's disease aren't

fully understood. But at a basic level, brain proteins fail to function normally, which disrupts the work of brain cells (neurons) and triggers a series of toxic events. Neurons are damaged, lose connections to each other and eventually destroy memory and other important mental functions. Memory loss and confusion are the main symptoms. No cure exists, but medication and management strategies may temporarily improve symptoms [4], [9].

Breast cancer is the most common cancer diagnosed in women, accounting for more than 1 in 10 new cancer diagnoses each year. It is the second most common cause of death from cancer among women in the world. Breast cancer evolves silently, and most disease is discovered on routine screening. Breast cancer has now overtaken lung cancer as the world's most commonly diagnosed cancer, according to statistics released by the International Agency for Research on Cancer (IARC) in December 2020. Identifying factors associated with an increased incidence of breast cancer development is important in general health screening for women [5], [10].

Predictive analytics automatically analyze databases using algorithms like single tree, decision tree, tree boost, support vector machine and neural network [3], [4], [7]-[9], [11]-[15].

The present study aims to analyze the datasets of selected diseases such as Alzheimer's and Breast cancer which is affecting a majority of population using the DTREG data mining tool to understand the algorithm which works best and the important factors to be studied for predicting these diseases. Thereby helping in the early prognosis and diagnosis of diseases like Alzheimer's and breast cancer.

METHODOLOGY

Two different diseases were selected for the study namely Alzheimer's and Breast cancer. Data mining tool used was DTREG. DTREG is a powerful application that can be easily installed on any Windows system. DTREG reads comma-separated value (CSV) data files that can be easily created from almost any data source. After creating the data file, just insert it into DTREG and DTREG complete all the work of creating decision trees, support vector machines, KMeans clustering, linear discriminant functions, linear regression or logistic regression models. Even complex analysis can be completed in minutes. DTREG can build classification trees and regression trees where the target variables are continuous, such as revenue or sales. The datasets chosen for the study were secondary data downloaded from Kaggle website, which is world's largest data science community with powerful tools and resources. This tool can be downloaded from the website *www.dtreg.com.* After installation of the tool, which is self-guided, the procedure to use the tool is explained below.

Procedure

A sample model was created by clicking on add project. Then the existing dataset was opened. On the project page details about the project were entered, i.e., title of project, input data file, data sub setting, character used to separate decimal point in input data file and character used to separate columns, notes about this project, etc. Then time series analysis was performed and the type of model to build was decided. Information about the variables was specified. The file was then saved. The file was then opened and Run analysis performed. A new report was displayed. The generated decision tree was viewed and the results interpreted.

RESULTS OF DATASETS USING DTREG

The parameters used in this Alzheimer's datasets were age, gender, socioeconomic status (SES), mini mental status examination (MMES), clinical dementia rating (CDR), estimated total intra cranial blood volume (eTIV), normalize whole brain volume (NWBV), atlas scaling factor (ASF) and the parameters used in breast cancer dataset include clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli and mitoses

Building a Model

There are five prediction models that were developed using classification technique and there were certain commonalities among all the models. All the models considered only 09 predictor variables and 01 target variable i.e., 'Group' for Alzheimer's and 'class' for Breast Cancer**.** The classification technique was used for analysis and the category weights were distributed over entire data file. The misclassification costs were equal (unitary) and the variable weights were also equal.

1. *Single tree Model:* Maximum splitting levels of single tree model is 10. Minimum size node to split is 10, whereas the minimum rows allowed in a node were 05. Maximum categories for continuous predictors were 1,000. Cross-validation method with ten folds was used for tree pruning and validation.

Model size for Alzheimer's - Maximum depth of the tree was 08. Total number of group splits was 27. The full tree has 15 terminal nodes. Minimum validation relative error occurred with 13 nodes. The relative error value was 0.2080 with a standard error of 0.0206 and the tree was pruned from 15 to 13 nodes.

Model size for Breast Cancer - The maximum depth of the tree was 08. Total number of group splits was 17. The full tree had 09 terminal nodes. The minimum validation relative error occurred with 08 nodes. The relative error value was 0.1212 with a standard error of 0.0119. The tree was pruned from 09 to 08 nodes

2. *Decision tree Model:* Maximum trees in Decision Tree Forest were 200. Maximum splitting levels was 50. Misclassification costs: equal (unitary). Minimum size node to split was 02 and maximum categories for continuous predictors were 100 for Alzheimer's and 1000 for Breast Cancer. Tree validation method was Out

of Bag (OOB).

Model size for Alzheimer's - The full forest had 200 trees. Three predictors (out of 09) were used for each split. Maximum depth of any tree in the forest was 15 and average number of group splits in each tree was 43.6.

Model size for Breast Cancer - Three predictors (out of 09) were used for each split. Maximum depth of any tree in the forest was 15. Average number of group splits in each tree was 27.4

Tree boost Model: Maximum trees in Tree Boost series were 400. Maximum splitting levels were 05. Minimum size node to split was 10. Maximum categories for continuous predictors were 1,000. Random sampling (20%) validation method was used. Tree pruning criterion was the minimum absolute error.

Model size for Alzheimer's - All 09 predictors were considered for each split. Maximum depth of any tree in the series was 05. Average number of group splits in each tree was 28.6. The minimum error with the training data and the test data occurred with 391 trees. Hence the tree series was pruned to 391.

Model size for Breast Cancer – In gradient tree boost model, all 09 predictors were considered for each split. Maximum depth of any tree in the series is 05. Average number of group splits in each tree was 9.4. The minimum error with the training data and the test data occurred with 311 trees and 104 trees respectively. Hence the tree series was pruned to 104 series.

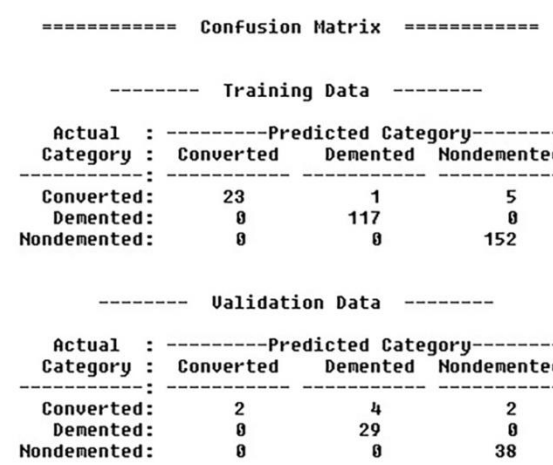
3. *Support Vector Machine Model:* The type of SVM model was C-SVM and the SVM kernel function was radial basis function. SVM grid and pattern searches found optimal values for the following parameters. The search criterion was to minimize total error.

For Alzheimer's dataset, the total number of points evaluated during search was 139 and the minimum error found by search was 0.069705. ($\varepsilon = 0.001$, $C = 29.7019381$, $\gamma = 5.74349177$). The number of support vectors used by the model was 193.

For Breast Cancer dataset, the total number of points evaluated during search was 148 and the minimum error found by search was 0.027818. ($\varepsilon = 0.001$, $C = 0.1$, $\gamma = 0.001$). The number of support vectors used by the model was 478.

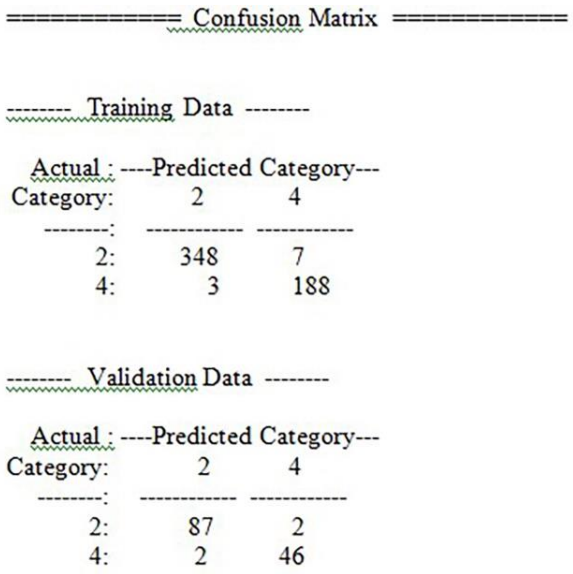
4. *Neural Network Model:* Confusion matrix of Alzheimer's and Breast Cancer is shown in Figure 1 and 2. Neural network technique in Alzheimer's dataset was used to predict whether the subject was converted, demented and non demented shown in Table I, II, III, IV. The probability values of occurrence of nondemented was found to be 0.5093834, demented was found to be 0.3914209 and the converted was found to be 0.0991957.

Neural network technique in Breast Cancer dataset was used to predict whether the subject belonged to class 2 or class 4 is given in Table V, VI. The probability values of occurrence of class 2 was found to be 0.6500732 and the class 4 was found to be 0.3499268.



```
============  Confusion Matrix  ============

--------  Training Data  --------

  Actual  : ---------Predicted Category--------
 Category :  Converted    Demented  Nondemented
-----------: ----------- ----------- -----------
Converted:        23           1           5
 Demented:         0         117           0
Nondemented:       0           0         152

--------  Validation Data  --------

  Actual  : ---------Predicted Category--------
 Category :  Converted    Demented  Nondemented
-----------: ----------- ----------- -----------
Converted:         2           4           2
 Demented:         0          29           0
Nondemented:       0           0          38
```

Fig. 1. Confusion Matrix of Alzheimer's Disease.



```
============ Confusion Matrix ============

-------- Training Data --------

Actual : ----Predicted Category---
Category:       2          4
--------: ----------- -----------
      2:        348          7
      4:          3        188

-------- Validation Data --------

Actual : ----Predicted Category---
Category:       2          4
--------: ----------- -----------
      2:         87          2
      4:          2         46
```

Fig. 2. Confusion Matrix of Breast Cancer

Table 1. Results for Converted Group of Alzheimer's Disease

	Single Tree Model		Decision Tree Model		Tree Boost Model		Support Vector Machine Model		Neural Network Model	
	Training data	*Validation data*	*Training data*	*Validation data*	*Training data*	*Validation data*	*Training data*	*Validation data*	*Training data*	*Validation data*
Total Records	373	373	373	373	298	75	373	373	373	373
Accuracy	93.57%	89.54%	91.69%	91.69%	97.99%	92.00%	99.73%	93.03%	100%	96.25%
Misclassification	93.03%	89.01%	91.15%	91.15%	97.99%	92.00%	99.73%	93.03%	100%	94.91%

Sensitivity	48.65%	24.32%	21.69%	21.62%	79.31%	25.00%	97.30%	45.95%	100%	75.68%
Specificity	98.51%	96.73%	99.40%	99.40%	100%	100%	100%	98.51%	100%	98.51%
Precision	78.26%	45.00%	80.00%	80.00%	100%	100%	100%	77.27%	100%	84.85%
Recall	48.65%	24.32%	21.62%	21.62%	79.31%	25.00%	97.30%	45.95%	100%	75.68%
Fmeasure	0.6000	0.3158	0.3404	0.3404	0.8846	0.4000	0.9863	0.5763	1.0000	0.8000

Table 2 Results For Demented Group of Alzheimer's Disease

	Single Tree Model		Decision Tree Model		Tree Boost Model		Support Vector Machine Model		Neural Network Model	
	Training data	*Validation data*	*Training data*	*Validation data*	*Training data*	*Validation data*	*Training data*	*Validation data*	*Training data*	*Validation data*
Total Records	373	373	373	373	298	75	373	373	373	373
Accuracy	95.71%	95.71%	95.71%	95.71%	99.66%	94.67%	100.00%	96.25%	100%	97.32%
Misclassification	93.03%	89.01%	91.15%	91.15%	97.99%	92.00%	99.73%	93.03%	100%	94.91%
Sensitivity	99.32%	99.32%	99.32%	99.32%	100.00%	100.00%	100.00%	97.95%	100%	95.21%
Specificity	93.39%	93.39%	93.39%	93.39%	99.45%	91.36%	100%	95.15%	100%	98.68%
Precision	90.63%	90.63%	90.63%	90.63%	99.15%	87.88%	100%	92.86%	100%	97.89%
Recall	98.63%	94.52%	99.32%	99.32%	100.00%	100.00%	100.00%	97.95%	95.21%	95.21%
Fmeasure	0.9505	0.9200	0.9477	0.9477	0.9957	0.9355	1.0000	0.9533	0.9653	0.9653

Table. 3 Results For Nondemented Group of Alzheimer's Disease

	Single Tree Model		Decision Tree Model		Tree Boost Model		Support Vector Machine Model		Neural Network Model	
	Training data	*Validation data*	*Training data*	*Validation data*	*Training data*	*Validation data*	*Training data*	*Validation data*	*Training data*	*Validation data*
Total Records	373	373	373	373	298	75	373	373	373	373
Accuracy	94.91%	94.91%	94.91%	94.91%	98.32%	97.33%	99.73%	96.51%	100%	96.25%
Misclassification	99.03%	89.01%	91.15%	91.15%	97.99%	92.00%	99.73%	93.03%	100%	94.91%
Sensitivity	98.42%	98.42%	98.42%	98.42%	100.00%	100.00%	100.00%	98.42%	100%	98.42%
Specificity	91.26%	91.26%	91.26%	91.26%	96.58%	94.59%	99.45%	94.54%	100%	93.99%
Precision	92.12%	92.12%	92.12%	92.12%	96.82%	95.00%	99.48%	94.92%	100%	94.42%
Recall	97.37%	97.37%	98.42%	98.42%	100.00%	100.00%	100.00%	98.42%	100%	98.42%
Fmeasure	0.9661	0.9512	0.9517	0.9517	0.9838	0.9744	0.9744	0.9664	1.0000	0.9639

Table 4. Lift and Gain of Alzheimer's Disease

Models	Training Data			Validation data		
	Converted	*Demented*	*Nondemented*	*Converted*	*Demented*	*Nondemented*
Single Tree	9.92%	39.14%	50.94%	9.92%	39.14%	50.94%
Decision Tree	9.92%	39.14%	50.94%	9.92%	39.14%	50.94%
Tree Boost	9.73%	39.26%	51.01%	10.67%	38.67%	50.67%
SVM	9.92%	39.14%	50.94%	9.92%	39.14%	50.94%
NN	9.92%	39.14%	50.94%	9.92%	39.14%	50.94%

Table 5. Results for Class 2 and Class 4 of Breast Cancer

	Single Tree Model		Decision Tree Model		Tree Boost Model		Support Vector Machine Model		Neural Network Model	
	Training data	*Validation data*	*Training data*	*Validation data*	*Training data*	*Validation data*	*Training data*	*Validation data*	*Training data*	*Validation data*
Total Records	683	683	683	683	546	137	683	683	683	683
Accuracy	97.07%	95.75%	97.36%	97.36%	98.17%	97.08%	97.22%	97.22%	100%	98.10%
Misclassification	97.07%	95.75%	97.36%	97.36%	98.17%	97.08%	97.22%	97.22%	100%	98.01%
Sensitivity	97.91%	93.72%	98.33%	98.33%	98.43%	95.83%	97.49%	97.91%	100%	97.91%
Specificity	96.62%	96.85%	96.85%	96.85%	98.03%	97.75%	97.07%	96.85%	100%	98.20%
Precision	93.98%	94.12%	94.38%	94.38%	96.41%	95.83%	94.49%	94.35%	100%	96.69%
Recall	97.91%	93.72%	98.33%	98.33%	98.43%	95.83%	97.49%	97.91%	100%	97.91%
Fmeasure	0.9590	0.9392	0.9631	0.9631	0.9741	0.9583	0.9608	0.9610	1.0000	0.9730
Probability error	0.0000	0.017731	0.013951	0.013951	0.163465	0.033374	0.023082	0.024865	0.000364	0.01693
AUROC	0.982689	0.968370	0.993012	0.993012	0.999189	0.986891	0.995373	0.995043	1.0000	0.993696

Table 6. Lift and Gain of Breast Cancer

Models	Training Data		Validation data	
	Class 2	*Class 4*	*Class 2*	*Class 4*
Single Tree	65.01%	34.99%	65.01%	34.99%
Decision Tree	65.01%	34.99%	65.01%	34.99%
Tree Boost	34.85%	35.98%	65.96%	35.04%
SVM	65.01%	34.99%	65.01%	34.99%
NN	65.01%	34.99%	65.01%	34.99%

Importance of Variable for Alzheimer's: The variable 'Clinical Dementia Rating (CDR)'and 'estimated total intra cranial blood volume(eTIV)' was the most important variable according to all the models. However, the variable 'Normalize whole brain volume(nWBU)' was the second important variable as per Decision Tree Forest model. However, the variable 'Age' was the next important variable as per the Tree Boost model.

Importance of Variable for Breast Cancer: It could be concluded from the study that the variable 'Uniformity of cell size' and 'uniformity of cell shape' was the most important variable. However, the variable 'Bare Nucleoli' was the next important variable as per Decision Tree Forest model, and the Tree Boost model.

CONCLUSION

The main purpose of the research work was to analyze the datasets of selected diseases (Alzheimer's and Breast cancer) using the data mining software DTREG. All the models built for predicting the category of Alzheimer's patients and the survivability of breast cancer patients showed similar results and performance. However, the Neural network model is marginally better than the others as all 09 predictors were considered for each spit. The experimental result of accuracy, sensitivity, area under ROC curve and lift-gain were also slightly better in the Neural network model. Thus, the Neural network model was effective and the best model for predicting Alzheimer's and the survivability of breast cancer. The significant predictors for Alzheimer's were total intracranial blood volume, clinical dementia rating and age and for breast cancer, uniformity of cell size, cell shape and clump thickness were significant.

Data mining and machine learning can thus be of very good help in deciding the line of treatment to be followed by extracting knowledge from such suitable databases. The study carried out is generally a clinical decision support system. In this study, predictions have been made for diagnosis and treatment. It helps decision makers with recommendations by using clinical data stack and patient-specific data especially created by internal medicine specialists. In the study, a comparison has been made between different algorithms that could be used for the component of inference mechanism which is the brain of the clinical decision support systems. Also, the important predictors have been identified. Many more datasets could be added to improve the prediction accuracy. Further, a greater number of diseases and the availability of more data mining tools could be explored.

ACKNOWLEDGMENT

Authors are highly thankful to the management of M S Ramaiah University of Applied Sciences, Bengaluru for allowing us to carry out our work at their premises.

REFERENCES

[1] S. T. Alanazi, M. Anbar, S. A. Ebad, S. Karuppayah, and H. A. Al-Ani, "Theory-based model and prediction analysis of information security compliance behavior in the Saudi healthcare sector," Symmetry, vol. 12, no. 9, pp. 1544, 2020.

[2] P. K. Sahoo, S. K. Mohapatra, and S. L. Wu, "Analyzing healthcare big data with prediction for future health condition," IEEE Access, vol. 4, pp. 9786-9799, 2016.

[3] V. Kumar, and M. L. Garg, "Predictive analytics: a review of trends and techniques," Int. J. Comput. Appl, vol. 182, no. 1, pp. 31-37, 2018.

[4] M. Bucholca, X. Ding, H. Wang, D. H. Glass, H. Wang, G. Prasad, L. P. Maguire, A. J. Bjourson, P. L. McClean, S. Todd, D. P. Finn, K. Wong-Lin, "A practical computerized decision support system for predicting the severity of Alzheimer's disease of an individual," Expert. Syst. Appl, vol. 15, no. 130, pp. 151-171, 2019.

[5] Z. K. Senturk, and R. Kara, "Breast cancer diagnosis via data mining: performance analysis of seven different algorithms," Computer Science and Engineering: An International Journal, vol. 4, no. 1, pp. 35-46, 2014.

[6] N. Sharma, and H. Om, "Framework for early detection and prevention of oral cancer using data mining," Int. J. adv. eng. Technol, vol. 4, no. 2, pp. 302-310, 2012.

[7] N. Sharma, and H. Om, "Comparing the performance of Data mining tools: WEKA and DTREG," Int. J. Sci. Eng. Res, vol. 5, no. 4, pp. 911- 918, 2014.

[8] D. Kaladhar, B. Chandana, and P. B. Kumar, "Predicting cancer survivability using classification algorithms," International Journal of Research and Reviews in Computer Science, vol. 2, no. 2, pp. 340-343, April 2011.

[9] A. Kumar, and T. R. Singh, "A new decision tree to solve the puzzle of Alzheimer's disease pathogenesis through standard diagnosis scoring system," Interdiscip Sci Comput Life Sci, vol. 9, no. 1, pp. 107-115, 2017.

[10] E. L. Cavalieri, and E. G. Rogan, "The etiology and prevention of breast cancer," Drug. Discov. Today. Dis. Mech, vol. 9, no. 1-2, pp. 55-69, 2013.

[11] V. Chaurasia, S. Pal, and B. B. Tiwari, "Prediction of benign and malignant breast cancer using data mining techniques," J. Algorithm. Comput. Technol, 2018.

[12] N. Sharma, and H. Om, "Data mining models for predicting oral cancer survivability," Netw. Model. Anal. Health. Inform. Bioinforma, vol. 2, no. 4, pp. 285-295, 2013.

[13] R. M. Rahman, and F. R. M. D. Hasan, "Using and comparing different decision tree classification techniques for mining ICDDR, B hospital surveillance data," Expert Syst Appl, vol. 38, pp. 11421–11436, 2011.

[14] S. Petruseva, V. Zileska-Pancovska, Z. Vahida, and A. B. Vejzovic, "Construction costs forecasting: comparison of the accuracy of linear regression and support vector machine models," Technical Gazette, vol. 24, no. 5, pp. 1431-1438, 2017.

[15] Obrzut, M. Kusy, A. Semczuk, M. Obrzut, and J. Kluska, "Prediction of 10-year overall survival in patients with operable cervical cancer using a probabilistic neural network," J. Cancer, vol. 10, no. 18, pp. 4189-4195, 2019.