Polyneuropathy disease forecast in the type 2 diabetes mellitus patients using data mining based approach

Nur Kuban Torun[1], Umman Tuğba Şimşek Gürsoy[2], Saadet Kader[3], M. Burak Oztop[4]
[1]Faculty of Economics and Administrative Sciences, Department of Business Administration/Quantitive Methods, Bilecik Seyh Edebali University, Bilecik,
[2]Faculty of Business Administration, Department of Quantitative Methods, Istanbul University, İstanbul,
[3]Department of Medical Biochemistry, Bilecik Public Health Laboratory, Bilecik,
[4]Department of General Surgery Bilecik Public Hospital, Bilecik, Turkey

This research consists of a part of the doctoral thesis research have being written by Nur Kuban Torun and conducted under the consultancy of Prof. Dr. U. Tugba Simsek Gursoy in the Department of Quantitive Methods, Istanbul University.

Abstract

Aim: The analytic quality of cardiac biomarkers were investigated consecutive six months by sigma metric method in our emergency laboratory. Total allowable error ratio (TEa%)'s of AAB, BV, RCPA, Ricos, and Rilibak were used for calculation. Sigma levels are compared and used to decide which TEa% is appropriate for our laboratory for more accurate results. Material and Method: Sigma levels were calculated for cardiac biomarkers which include Troponin I (cTnI), Troponin T (cTnT), CKMB mass, Myoglobin (Mb) and NT-proBNP in our emergency laboratory department between December 2017 and May 2018. The internal quality control (IQC) and external quality control (EQC) assessment results and TEa%'s of AAB, BV, RCPA, Ricos, and Rilibak were used to calculate sigma metrics. The sigma metrics for tests were calculated by "Sigma = (TEa% – Bias%) / CV%" formula. Results: Considering different TEa% 's, it is evaluated that CKMB mass sigma level is at the "world-class quality". On the contrary, cTnT sigma level is found to be at the level of "poor quality". For AAB, BV, RCPA, Ricos and Rilibak, different sigma levels are observed. Discussion: Due to using different TEa%'s for each test, different sigma levels were determined. On the other hand, because of the "poor quality" level of cTnT sigma value, decision is taken for the improvement of cTnT in our laboratory. In addition, it is observed that there is no specified TEa% for whole blood samples. Therefore, it is concluded that, for more accurate and consistent evaluations, specified matrix of TEa% values are required for whole blood samples.

Keywords

Data Mining; C4.5; Random Forest Tree; Polyneuropathy Disease; Diabetes Mellitus

DOI: 10.4328/JCAM.6000 Received: 08.08.2018 Accepted: 07.09.2018 Published Online: 17.09.2018
Corresponding Author: Nur Kuban Torun, Faculty of Economics and Administrative Sciences, Department of Business Administration/Quantitive Methods, Bilecik Seyh Edebali University, Bilecik, Turkey. GSM: 905534811839 E-Mail: akdemir.kuban@gmail.com
ORCID ID: 0000-0002-9115-5838

Introduction

Nowadays, due to the increasingly growing IT technologies, every single data of a patient is recording and storing in databases. These bulk datum comprise utility knowledge for both profit and non-profit organizations. However, it's a big issue that transforming datum into the reliable and significant knowledge to meet the practitioners' expectation. Therefore, there is a concept named data mining [1]. The main manner of data mining is mining the datum to create the intelligible and utility knowledge by analyzing the hidden patterns and relationships between given attributes [2]. Through the growing usage of Internet technologies in almost all fields, data mining has increased attention and utilization in both profit and non-profit companies. In this manner, there has been a new era in healthcare and medical via data mining modeling [3]. The research about data mining shows that it will help healthcare and medicine field to redound the success in diagnosis, patient services, prevent of the prevalence of chronic diseases, the efficiency of the health budget, cut of corruptions and etc. [4]. Data mining in healthcare and medicine is called health information that is a mixture of bio-information, clinical information, public health information, and neuro-information. These fields contribute to health information to gain datum and health information analyzes these datum to create significant information and store to use in the future [5].

Healthcare data mining is based on the datum of clinical records and used to determine the risk factors and prevalence [6]. Accordingly, data mining needs both qualitative and quantitative datum to use in data mining algorithms. These algorithms are clustering and making classification to the given datum to predict or estimate a situation. The classification in data mining is being used for prevention of mortality risk related to the diseases of cancer, diabetes mellitus and cardiovascular [7]. On the other hand, patients' printed documents and e-records are vitally important to gain information about the correlations between disease and its reasons. For this purpose, text mining and clustering algorithms are being used to put forward these relationships [8]. Herewith, the algorithms of data mining being used in healthcare can be divided into 2 subgroups which are classification algorithms and clustering algorithms. Classification algorithms include decision tree, k-nearest neighbor, neural networks, support machine, Naive Bayes and logistic regression. The most used clustering algorithm is K-means clusters. These algorithms are applied in specific software to obtain knowledge [9].

Post-modern era evokes the over-consumption of people. Consequently, the chronic disease such as diabetes mellitus is also increasing, and it is a major problem that will be ended by mortality all over the world [10]. There are too many people have diabetes mellitus all over the world [11]. Thence, it is important to estimate the prevalence of diabetes mellitus via data mining. In order to obtain the data on diabetes mellitus, data warehouse including e-records, clinical data, inspection values, personal information, heritability and interviews with patients is substantial [12].

Material and Method

There are 7 variables related to diabetes mellitus to predict the diabetic polyneuropathy. The variables are Diabetic Polyneuropathy, Gender, Glycated Haemoglobin (HbA1c), Creatinine, Total Cholesterol, High-Density Lipoprotein (HDL), and Low-Density Lipoprotein (LDL). Datum was obtained from local health institution and includes 2907 type 2 diabetes mellitus patients. The variables are following:

a- Diabetic Polyneuropathy:

Diabetic neuropathies are a heterogeneous group of pathological manifestations with the potential to affect every organ, with clinical implications such as organ dysfunction, which leads to low-quality life and increased morbidity. DPN is defined as peripheral nerve dysfunction with positive and negative symptoms. Risk factors include age, male gender, duration of diabetes, uncontrolled glycemia, height, overweight and obesity, and insulin treatment [13].

b- Gender:

There is increasing evidence that sex and gender differences are important in epidemiology, pathophysiology, treatment, and outcomes in many diseases, but they appear to be particularly relevant for noncommunicable diseases. Sex differences describe biology-linked differences between women and men, which are caused by differences in sex chromosomes, sex-specific gene expression of autosomes, sex hormones, and their effects on organ systems. Both biological and psychosocial factors are responsible for sex and gender differences in diabetes risk and outcome [10].

c- Glycated Haemoglobin (HbA1c):

HbA1c is a blood test that measures the average blood glucose level over the previous 3–4 months [14].

d- Creatinine:

Creatinine is a waste product of muscle metabolism that is normally removed by the kidneys. The presence of excess creatinine is an indication of increased muscle breakdown or a disruption of kidney function [14].

e- Total Cholesterol:

Total cholesterol is measured in terms of milligrams (mg) per deciliter (dL) of blood. A milligram is equal to one-thousandth of a gram. A deciliter is equal to one-tenth of a liter. Desirable levels are below 200 mg/dL. Borderline high levels are 200–239 mg/dL.

High levels are 240 mg/dL and above [14].

f- Low-Density Lipoprotein (LDL):

LDL is referred to as bad cholesterol because excess quantities of LDL contribute to plaque buildup in the arteries. Optimal levels are below 100 mg/dL. Near optimal is between 100 and 129 mg/dL. Borderline high level is between 130 and 159 mg/dL. High level is between 160 and 189 mg/dL. Very high level is 190 mg/dL and above [14].

g- High-Density Lipoprotein (HDL):

HDL is referred to as a good cholesterol because it carries unneeded cholesterol back to the liver for processing and does

not contribute to plaque buildup. Bad levels are below 40 mg/dL. Better levels are between 40 and 59 mg/dL. Best levels are 60 mg/dL and above [14].

Statistical Analysis

All variables were subjected to the data mining algorithms that are C4.5 decision tree and random forest algorithm to estimate the diabetic polyneuropathy. Accordingly, the software of R programming was also used to apply these algorithms and findings were noted.

a-C4.5. Decision Tree

A decision tree is a classifier expressed as recursive partition of the instance space. The decision tree consists of nodes that branch within a rooted tree. It starts with a root at the top that has no incoming edges. A node with outgoing edges is called an internal node, and all the other nodes are called leaves, also known as decision nodes. Each leaf is assigned to one class representing the majority target value at that node [15].

b- Random Forest Tree

The RF algorithm, which is widely used for classification in bioinformatics, builds *nTree* (a parameter) Random Trees (RT) during its training phase. This involves randomizing the training set in two ways for each RT: First, the training set is resampled with replacement, maintaining the original size of the dataset. As a second source of randomness for building an RT, the search for the best feature to split the set of instances at each RT node considers a randomly chosen feature subset of size mtry (a parameter), typically much smaller than the original feature set's size. The instances at the current node are then split into two subsets according to a condition based on the values of the selected feature, creating two child nodes. This split aims to increase the similarity of classes within each instance subset and to decrease class similarity across the subsets. Next, the algorithm recurses in each instance subset until a stopping criterion is met [16].

Results

Due to the processing of data mining, the following steps are applying [17]:

Pre-processing:

In the given dataset, there were numbers of noisy datum that affect the modelling process. These noisy data are occurred by incorrect non-numerical columns. Accordingly, handle with these missing values, the clearance step was applied and some of the data were removed from the dataset.

Table 1. Transformation of HbA1c Dataset

HbA1c	Accepted HbA1c Intervals	The Number of Patient out of 2907	Percentage (%)
Normal	x <5,7	166 patient	6
Impaired fasting glucose	[5,7-6,4]	641 patient	22
Diabetes Mellitus Type 2	x>6,4	2100 patient	72

Table 2. Transformation of Creatinine Dataset

Creatinine	Accepted Creatinine intervals for Male	Accepted Creatinine intervals for Female	The Number Of Patient out of 2907	Percentage (%)
Low	x<0,6	x<0,5	42	2
Normal	[0,6-1,2]	[0,5-1,1]	2492	86
High	x>1,2	x>1,1	373	12

Table 3. Transformation of Total Cholesterol Dataset

Total cholesterol	Accepted total cholesterol intervals	The Number Of Patient out of 2907	Percentage (%)
Desirable levels	x<200	1346	46
Borderline levels	[200-240]	893	31
High levels	x>240	663	23

Table 4. Transformation of HDL Dataset

HDL Cholesterol	Accepted HDL cholesterol intervals for Male	Accepted HDL cholesterol intervals for Female	The Number Of Patient out of 2907	Percentage (%)
Bad levels	x<40	x<50	1257	43
Better levels	[40-60]	[50-60]	1256	42
Best levels	x>60	x>60	394	19

Table 5. Transformation of LDL Dataset

LDL Cholesterol	Accepted LDL cholesterol intervals	The Number Of Patient out of 2907	Percentage (%)
Optimal levels	x<100	860	30
Near optimal levels	[100-129]	888	31
Borderline high levels	[130-159]	681	23
High levels	[160-189]	350	12
Very high levels	x>189	128	4

Transformation of the Datum:

Due to the nature of using algorithms of datamining, dataset numbers were transformed into percentage values shown in Table 1, Table 2, Table 3, Table 4 and Table 5. The HbA1c da-

Table 6. Summary of Datas in the R Programming

"Diabetic polyneuropathy"							
Age	Gender	HbA1c	Creatinine	Total cholesterol	HDL	LDL	Diabetic Polyneuropathy
Min. :16,00000	1:1167	Min. : 6,00000	Min. : 2,00000	Min. :23,00000	Min. :19,0000	Min. : 4,00000	Yes: 222
1st Qu.:52,00000	2:1740	1st Qu.:22,00000	1st Qu.:86,00000	1st Qu.:31,00000	1st Qu.:42,0000	1st Qu.:23,00000	No: 2685
Median :60,00000		Median :72,00000	Median :86,00000	Median :31,00000	Median :42,0000	Median :30,00000	
Mean :59,66598		Mean :57,20605	Mean :75,29137	Mean :36,10698	Mean :39,3151	Mean :25,35363	
3rd Qu.:68,00000		3rd Qu.:72,00000	3rd Qu.:86,00000	3rd Qu.:46,00000	Mean :39,3151	3rd Qu.:31,00000	
Max. :91,00000		Max. :72,00000	Max. :86,00000	Max. :46,00000	Max. :43,0000	Max. :31,00000	

Table 7. The Accuracy of C4.5 Algorithm

	Estimation	
Acceptance	Available	Unavailable
Available	1	43
Unavailable	3	534

"Accuracy= 0,920826161790017"

Table 8. The Accuracy of Random Forest Tree Algorithm

	Estimation	
Acceptance	Available	Unavailable
Available	1	43
Unavailable	2	535

"Accuracy 0,922547332185886"

taset is divided into 3 sub-groups. The first one is the normal group, the second one is the impaired fasting glucose and the third group is the diabetes mellitus type 2. The accepted values for the normal group is below 5,7 mmol/L, the impaired fasting glucose is 5,7-6,4 mmol/L, and the diabetes mellitus type 2 is 6,4 mmol/L and above. Accordingly, normal sub-group includes 166 patients was transformed into 6%, impaired fasting glucose sub-group includes 641 patients was transformed into 22%, and diabetes mellitus type 2 sub-group includes 2100 patients was transformed into 72%.

The Creatinine dataset was divided into 3 sub-groups. The first one is the low Creatinine, the second one is the normal Creatinine and the third one is the high Creatinine. The accepted values for the low Creatinine is below 0,6 mg/dL for male and below 0,5 mg/dL for female, the normal Creatinine is 0,6-1,2 mg/dL for male and 0,5-1,1 mg/dL for female and the high Creatinine is 1,2 mg/dL and above for male and 1,1 mg/dL and above for female. The percentage of the low Creatinine is 2%, the normal Creatinine is 86%, and the high Creatinine is 12%.

The Total Cholesterol dataset was divided into 3 sub-groups. The first one is the desirable levels, the second one is the borderline levels and the third one is the high levels. The accepted values for the desirable levels are below 200 mg/dL, the borderline levels are 200-240 mg/dL and the high levels are 240 mg/dL and above. The percentages of the desirable levels are 46%, of the borderline levels are 31%, and of the high levels are 23%.

The HDL data set was divided into 3 sub-groups. The first one is the bad levels, the second one is better levels and the third one is the best levels. The accepted values for the bad levels are below 40 mg/dL for male and below 50 mg/dL for female, the better levels are 40-60 mg/dL for male and 50-60 mg/dL for female and the best levels are 60 mg/dL and above for male and 60 mg/dL and above for female. The percentages of the bad levels are 43%, the better levels are 42% and the best levels are 19%.

The LDL dataset was divided into 5 sub-groups. The first one is the optimal levels, the second one is the near optimal levels, the third one is the borderline high levels, the fourth one is the high levels and the fifth one is the very high levels. The accepted values for the optimal levels are below 100 mg/dL, the near optimal levels are 100-129 mg/dL, the borderline high levels are 130-159 mg/dL, the high levels are 160-189 mg/dL, and the very high levels are 189 mg/dL and above. The percentages of the optimal levels are 30%, the near optimal levels are 31%, the borderline high levels are 23%, the high levels are 12%, and the very high levels are 4%.

Processing

Data were analyzed with R programming based on the classification algorithms and shown in Table 6. In the R programming, the random start point was obtained by "*set.seed()*" function code. Afterward, the dataset was divided into the training set and the test set. The training set is a learning process to ascertain the algorithms and the test set is for testing the accuracy of this algorithm. In addition to accuracy, for evaluating the performance of the algorithms, the correspondence matrix was used.

Data Mining Process for C4.5.

The results of data mining with C4.5 algorithm are shown in Figure 1 and the accuracy rate is shown in Table 7. Due to the results, the accuracy value is 92%. Accordingly, the codes are:

- J48 pruned tree
- Creatinine <= 2
- HbA1c <= 6: available (2.0)
- HbA1c> 6
- LDL <= 12
- LDL <= 4: unavailable (3.0)
- LDL > 4: available (6.0/1.0)
- LDL > 12: unavailable (24.0/1.0)
- Creatinine > 2: unavailable (2291.0/170.0)
- Number of Leaves : 5
- Size of the tree : 9

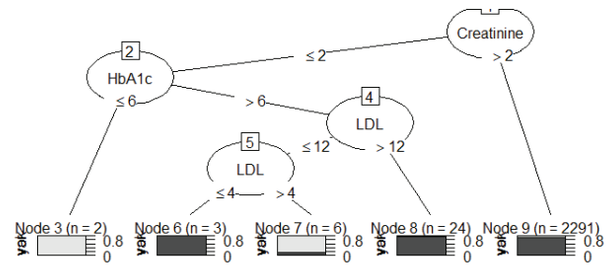

Figure 1. Datamining with C4.5.

According to the codes, the rules of C4.5 algorithm data mining are following:

Rule 1: If Creatinine <=2 and if HbA1c <= 6 then diabetic polyneuropathy: available.

Rule 2: If Creatinine <=2 and if HbA1c > 6 and if LDL<=12 and if LDL<=4 then diabetic polyneuropathy: unavailable.

Rule 3: If Creatinine <=2 and if HbA1c>6 and if LDL<=12 and if LDL>4 then diabetic polyneuropathy is available.

Rule 4: If Creatinine <=2 and if HbA1c>6 and if LDL>12 then diabetic polyneuropathy: unavailable.

Rule 5: If Creatinine >2, then diabetic polyneuropathy: unavailable.

Data Mining Process for Random Forest Tree

In the Random Forrest Tree algorithm, 500 trees were found. The number of variables tired at each split is 2. The correspondence matrix and the accuracy are shown in Table 8. The accuracy value is %92.

Discussion

According to the results, there are 5 rules for the availability of polyneuropathy disease. Due to the rule 1, if Creatinine ≤ 2 and if HbA1c ≤ 6 then polyneuropathy disease is available. In this condition, we should look at the Creatinine table. In this expression, 2 is a percentage and in the real data set, it refers to the 0,6 mg/dL for male and 0,5 mg/dL for female. Likewise, the score of HbA1c refers to 5,7 mmol/L. Hereby, the rule 1 is:

If Creatinine ≤ 0,6 mg/dL and HbA1c≤5,7 mmol/L then polyneuropathy disease is available for male.

If Creatinine ≤ 0,5 mg/dL and HbA1c≤5,7 mmol/L then polyneuropathy disease is available for female.

Considering the rule 2, if we peruse the related tables, we acquire the following formula:

If Creatinine ≤ 0,5 mg/dL and HbA1c >5,7 mmol/L and 160 ≤ LDL ≤ 189 mg/dL and LDL ≤ 189 mg/dL then polyneuropathy disease is unavailable.

Considering the rule 3, the formula is:

If Creatinine ≤ 0,5 mg/dL and HbA1c > 5,7 mmol/L and 160 ≤ LDL ≤ 189 mg/dL and LDL>189 mg/dL, then polyneuropathy disease is unavailable.

Considering the rule 4, the formula is:

If Creatinine ≤ 0,5 mg/dL and HbA1c >5,7 mmol/L and 160 ≤ LDL ≤ 189 mg/dL, then polyneuropathy disease is unavailable.

Considering the rule 5, the formula is:

If Creatinine >0,5 mg/dL then polyneuropathy disease is unavailable.

The conclusion of these rules is Creatinine, LDL and HbA1c are the primary three determinative factors on polyneuropathy disease. However, through the given dataset, Gender, HDL and Total Cholesterol have no significant relationships with the prediction of polyneuropathy disease.

On the other hand, one of the foremost results of this research is the accuracy of C4.5 classification algorithm. Previous significant studies on diabetes mellitus put forth the accuracy value of their C4.5 classification algorithm. Lakshmi and Kumar (18) used C4.5 classification algorithm to predict the diabetes mellitus. In this research, the accuracy value of the C4.5 was found at 72%. Another research is Radha and Srinivasan's study (19) used C4.5 algorithm to clinical data set to predict the diabetes mellitus. In this research, the accuracy value of C4.5 algorithm was found at 86%. Furthermore, Devi and Shyla (20) consulted to C4.5 algorithm in their research to predict the diabetes mellitus. In this study, C4.5 algorithm's accuracy was found at 86%. In our study, the accuracy value was found 0,920826161790017. Comparing with these studies, C4.5 algorithm is running with a high accuracy value of 92%. The score elucidates that the model can estimate 2154 instances correctly in 2326 classified instances. Moreover, the random forest creates 500 trees and it has almost the same accuracy value with C4.5. The accuracy for the Random Forest Tree is 0,922547332185886 and it shows that the model is in high accuracy in the value of 92%.

Consequently, to determine the availability of polyneuropathy disease through the given data mining algorithms, researchers may consider the Creatinine, LDL and HbA1c scores. However, the data set variables are the main limitation of this study. The model estimates the polyneuropathy disease with 6 variables. In the future, the model would be tested with varied variables and the results would be compared with the current outcomes. Another limitation is the generalization problem. The results are only valid for the discussed instance, and outcomes could be changeable with other population.

Scientific Responsibility Statement

The authors declare that they are responsible for the article's scientific content including study design, data collection, analysis and interpretation, writing, some of the main line, or all of the preparation and scientific review of the contents and approval of the final version of the article.

Animal and human rights statement

All procedures performed in this study were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. No animal or human studies were carried out by the authors for this article.

Funding: None

Conflict of interest

None of the authors received any type of financial support that could be considered potential conflict of interest regarding the manuscript or its submission.

References

1. Jiawei H, Kamber M. Data mining: concepts and techniques. CA: Morgan Kaufmann; 2006. p.5.
2. Hand D, Mannila H, Smyth P. Principles of data mining. USA: The MIT Press; 2001. p.6.
3. Earley S. The promise of healthcare analytics. IT Professional. 2015; 17(2): 7–9.
4. Koyuncugil A.S, Özgülbaş N. Veri madenciliği: tıp ve sağlık hizmetlerinde kullanımı ve uygulamaları. Bilişim Teknolojileri Dergisi. 2009; 2(2): 21-32.
5. Herland M, Khoshgoftaar T. M, Wald R. A review of data mining using big data in health informatics. Journal of Big Data. 2014; 1(1): 1-35.
6. Ressing M, Blettner M, Klug S. Data analysis of epidemiological studies. Dtsch Arztebl Int. 2010: 107(11); 187–92.
7. Sharma R, Singh S. N, Khatri S. Medical data mining using different classification and clustering techniques: a critical survey. Computational Intelligence & Communication Technology (CICT). 2016: 687–91.
8. Raja U, Mitchell T, Day T, Hardin J. M. Text mining in healthcare: applications and opportunities. Journal of Healthcare Information Management. 2008: 22(3): 52–6.
9. Bramer, M. Principles of data mining. UK: Springer; 2007. p.5-8.
10. Kautzky-Willer A, Harreiter J, Pacini G. Sex and gender differences in risk, pathophysiology and complications of type 2 diabetes mellitus. Endocr. Rev. 2016; 37(3): 278-316.
11. Shaw J.E, Sicree R.A, Zimmet P.Z. Global estimates of the prevalence of diabetes for 2010 and 2030. Diabetes Research and Clinical Practice. 2010; 87: 4-14.
12. Kob H. C, Tan G. Data mining applications in healthcare. Journal of Healtcare Information Management. 2011; 19(2): 64-72.
13. Román-Pintos L. M, Villegas-Rivera G, Cardona-Munoz E.G, Rodríguez-Carrizalez A. D, Moreno-Ulloa A, Rubin N, et al. Diabetic polyneuropathy in type 2 diabetes mellitus: inflammation, oxidative stress, and mitochondrial function. J. Diabetes Res. 2016: 1-16.
14. Ehrlich A, Schroeder C.L. Medical terminology. NY: Delmar Cangage; 2009. p.144.
15. Maimon O, Rokach L. Datamining and knowledge discovery handbook. Berlin: Springer; 2010. p.151.
16. Fabis F, Doherty A, Palmer D, Magalhaes J.P.D, Freitas A.A. A new approach for interpreting random forest models and its application to the biology of ageing. Bioinformatics. 2018; 34(14): 2449-56.
17. Larose D. T. Discovering knowledge in data: an introduction to data mining. Canada: Wiley-Interscience; 2005. p.27.
18. Lakshmi K. R. Utilization of data mining techniques for prediction and diagnosis of tuberculosis disease survivability. International journal of modern education and computer science. 2013: 5(8): 8–17.
19. Radha P, Srinivasan B. Predicting diabetes by cosequencing the various data mining classification techniques. International journal of ınnovative science, engi-

neering & technology. 2014: 1(6): 334–9.
20. Devi M. R, Shyla J. M. Analysis of various data mining techniques to predict diabetes mellitus. International journal of applied engineering research. 2016: 11(1): 727–30.

How to cite this article:
Torun NK, Şimşek Gürsoy UT, Kader S, Oztop MB. Polyneuropathy disease forecast in the type 2 diabetes mellitus patients using data mining based approach. J Clin Anal Med 2018; DOI: 10.4328/JCAM.6000.