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

Nur Kuban Torun, Umman Tugba Simsek Gursoy, Saadet Kader, M. Burak Oztop

1Faculty of Economics and Administrative Sciences, Department of Business Administration/Quantitive Methods, Bilecik Seyh Edebali University, Bilecik,  
2Faculty of Business Administration, Department of Quantitative Methods, Istanbul University, Istanbul,  
3Department of Medical Biochemistry, Bilecik Public Health Laboratory, Bilecik,  
4Department of General Surgery Bilecik Public Hospital, Bilecik, Turkey

**Abstract**

Aim: The aim of this research was to predict the availability of polyneuropathy disease of the type 2 diabetes mellitus patient through data mining algorithms.

Material and Method: The dataset was obtained from the Bilecik Public Hospital and the instance number is 2907. Models were created with two different classification data mining algorithms. The data set includes Gender, Glycated Haemoglobin (HbA1c), Creatinine, Total Cholesterol, Low-Density Lipoprotein (LDL) and High-Density Lipoprotein (HDL). Numerical data were transformed into interval forms and the percentiles were calculated for each interval. Results: Data analysis and performance evaluation were performed with R, RStudio. Random Forest Tree was found as the best algorithm for polyneuropathy disease prediction (the accuracy = 0.922547332185886). The accuracy of the C4.5 was found 0.920826161790017. The percentages of the normal levels HbA1c are 6%, the impaired fasting glucose levels are 22% and the diabetes mellitus type 2 levels are %72. The percentage of the low Creatinine is 2%, the normal Creatinine is %86 and the high Creatinine is 12%. The percentage of the desirable levels Total Cholesterol is 46%, the percentage of the borderline levels of Total Cholesterol is 31% and the percentage of high levels of Total Cholesterol is 23%. The percentage of the near optimal levels of LDL is 31%, the percentage of the borderline high levels of LDL is 23%, the percentage of the high levels of LDL is 12% and the percentage of the very high levels of LDL is 4%. The percentage of the bad levels of HDL is 43%, the percentage of the better levels of HDL is 42% and the percentage of the best levels HDL is 19%. This model indicated that Creatinine, LDL and HbA1c are the primary three determinative factors on polyneuropathy disease. Furthermore, the model created the following 5 rules. Rule 1: 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. Rule 2: 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. Rule 3: 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. Rule 4: 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. Rule 5: If Creatinine >0,5 mg/dL then polyneuropathy disease is unavailable.

Discussion: The results show that HDL, Gender and Total Cholesterol have no significant effect on the polyneuropathy disease in this model. To determine the availability of polyneuropathy disease through the given data mining algorithms, researchers may consider the Creatinine, LDL and HbA1c scores.

**Keywords**

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

DOI:10.4528/ACAM.6000  Received: 08.08.2018  Accepted: 07.09.2018  Published Online: 17.09.2018  Printed: 01.07.2019  Am Clin Anal Med 2019;10(4): 485-90  Corresponding Author: Nur Kuban Torun, Faculty of Economics and Administrative Sciences, Department of Business Administration/Quantitive Methods, Bilecik Seyh Edebali University, Bilecik, Turkey. GSM: 905534811859  E-Mail: akdemir.kuban@gmail.com  ORCID ID: https://orcid.org/ 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 rebond 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 unnecessary cholesterol back to the liver for processing and does...
Polynuearopathy disease forecast using data mining based approach

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.

\textbf{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].

\textbf{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].

\textbf{Results}
Due to the processing of data mining, the following steps are applying [17]:

\textbf{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.

\begin{table}[h]
\centering
\caption{Table 1. Transformation of HbA1c Dataset}
\begin{tabular}{|c|c|c|}
\hline
HbA1c & Accepted HbA1c & The Number Of Patient out of 2907 \\
\hline
Normal & x < 7 & 166 patient \\
Impaired fasting glucose & [5.7-6.4] & 641 patient \\
Diabetes Mellitus Type 2 & x > 6.4 & 2100 patient \\
\hline
\end{tabular}
\end{table}

\begin{table}[h]
\centering
\caption{Table 2. Transformation of Creatinine Dataset}
\begin{tabular}{|c|c|c|}
\hline
Creatinine & Accepted Creatinine intervals for Male & Accepted Creatinine intervals for Female \\
\hline
Low & x < 0.6 & x < 0.5 \\
Normal & [0.6-1.2] & [0.5-1.1] \\
High & x > 1.2 & x > 1.1 \\
\hline
\end{tabular}
\end{table}

\begin{table}[h]
\centering
\caption{Table 3. Transformation of Total Cholesterol Dataset}
\begin{tabular}{|c|c|c|}
\hline
Total cholesterol & Accepted total cholesterol intervals & The Number Of Patient out of 2907 \\
\hline
Desirable levels & x < 200 & 1546 \\
Borderline levels & [200-240] & 893 \\
High levels & x > 240 & 663 \\
\hline
\end{tabular}
\end{table}

\begin{table}[h]
\centering
\caption{Table 4. Transformation of HDL Dataset}
\begin{tabular}{|c|c|c|}
\hline
HDL Cholesterol & Accepted HDL cholesterol intervals for Male & Accepted HDL cholesterol intervals for Female \\
\hline
Bad levels & x < 40 & x < 50 \\
Better levels & [40-60] & [50-60] \\
Best levels & x > 60 & x > 60 \\
\hline
\end{tabular}
\end{table}

\begin{table}[h]
\centering
\caption{Table 5. Transformation of LDL Dataset}
\begin{tabular}{|c|c|c|}
\hline
LDL Cholesterol & Accepted LDL cholesterol intervals & The Number Of Patient out of 2907 \\
\hline
Optimal levels & x < 100 & 860 \\
Near optimal levels & [100-129] & 888 \\
Borderline high levels & [130-159] & 681 \\
High levels & [160-189] & 350 \\
Very high levels & x > 189 & 128 \\
\hline
\end{tabular}
\end{table}

\begin{table}[h]
\centering
\caption{Table 6. Summary of Datas in the R Programming}
\begin{tabular}{|c|c|c|c|c|c|c|c|}
\hline
\textbf{Age} & \textbf{Gender} & \textbf{HbA1c} & \textbf{Creatinine} & \textbf{Total cholesterol} & \textbf{HDL} & \textbf{LDL} & \textbf{Diabetic Polyneuropathy} \\
\hline
1st Qu. & 52.00000 & 2:1740 & 1st Qu. :22.00000 & 1st Qu. :86.00000 & 1st Qu. :31.00000 & 1st Qu. :42.00000 & 1st Qu. :25.00000 \\
Median & 60.00000 & 7:20000 & Median : 86.00000 & Median :31.00000 & Median :42.00000 & Median :30.00000 & \\
Mean & 59.66598 & 57.20050 & Mean :75.29137 & Mean :36.10698 & Mean :39.3151 & Mean :25.35363 & \\
3rd Qu. & 68.00000 & 72.00000 & 3rd Qu. :86.00000 & 3rd Qu. :46.00000 & 3rd Qu. :31.00000 & & \\
Max. & 91.00000 & 72.00000 & Max. :86.00000 & Max. :46.00000 & Max. :43.00000 & Max. :31.00000 & \\
\hline
\end{tabular}
\end{table}

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-
The Annals of Clinical and Analytical Medicine

Polyneuropathy disease forecast using data mining based approach

The LDL dataset was divided into 5 sub-groups. The first one is the best levels, the second one is the near optimal levels, the third one is the borderline high levels, the fourth one is the optimal levels and the fifth one is the very high levels. The accepted values for the low LDL is below 100 mg/dL, the borderline low levels are 100-129 mg/dL, the near optimal levels are 130-159 mg/dL, the high levels are 160-189 mg/dL, and the very high levels are 190 mg/dL and above. The percentages of the desirable levels are 20%, the near desirable levels are 29%, the bad levels are 43%, the better levels are 42% and the best levels are 60 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 low HDL, the second one is the normal HDL and the third one is the high HDL. The accepted values for the low HDL is below 50 mg/dL for male and below 60 mg/dL for female, the normal HDL is 50-129 mg/dL for male and 60-129 mg/dL for female and the high HDL is 130 mg/dL and above for male and 150 mg/dL and above for female. The percentage of the low HDL is 2%, the normal HDL is 86% and the high HDL 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 low HDL, the second one is the normal HDL and the third one is the high HDL. The accepted values for the low HDL is below 50 mg/dL for male and below 60 mg/dL for female, the normal HDL is 50-129 mg/dL for male and 60-129 mg/dL for female and the high HDL is 130 mg/dL and above for male and 150 mg/dL and above for female. The percentage of the low HDL is 2%, the normal HDL is 86% and the high HDL is 12%.

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 1 mg/dL for male and below 1,2 mg/dL for female, the normal Creatinine is 1,2-1,4 mg/dL for male and 1,1-1,2 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%.

Table 7. The Accuracy of C4.5 Algorithm

<table>
<thead>
<tr>
<th>Acceptance</th>
<th>Available</th>
<th>Unavailable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Available</td>
<td>1</td>
<td>43</td>
</tr>
<tr>
<td>Unavailable</td>
<td>3</td>
<td>534</td>
</tr>
</tbody>
</table>

*Accuracy= 0.920826161790017*

Table 8. The Accuracy of Random Forest Tree Algorithm

<table>
<thead>
<tr>
<th>Acceptance</th>
<th>Available</th>
<th>Unavailable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Available</td>
<td>1</td>
<td>43</td>
</tr>
<tr>
<td>Unavailable</td>
<td>2</td>
<td>535</td>
</tr>
</tbody>
</table>

*Accuracy= 0.922547332185886*

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

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 neuropathy: available.
Rule 2: If Creatinine <=2 and if HbA1c > 6 and if LDL<=12 and if LDL<=4 then diabetic neuropathy: unavailable.
Rule 3: If Creatinine <=2 and if HbA1c>6 and if LDL<=12 and if LDL>4 then diabetic neuropathy is available.
Rule 4: If Creatinine <=2 and if HbA1c>6 and if LDL>12 then diabetic neuropathy: unavailable.
Rule 5: If Creatinine >2, then diabetic neuropathy: 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, 0.9208261679790017. 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.

This research consists of a part of the doctoral thesis research 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.

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.

How to cite this article: