Leveraging Pharmacy Medical Records To Predict Diabetes Using A Random Forest & Artificial Neural Network

Stephen Lavery[1] and Jeremy Debattista[2]

[1] National College of Ireland, Dublin, Ireland, laverys@tcd.ie
[2] Trinity College Dublin, Dublin, Ireland, debattij@tcd.ie

Abstract. Diabetes is a disease that affects millions of people around the world. As early diagnosis of the disease is critical, predictive models have been designed in order to classify undiagnosed patients. This study proposes the use of pharmacy medical records for tackling this problem, which have previously remained an untapped resource. As the volume of pharmacy data is greater than that of clinical datasets, models developed using these records could be deployed on a larger scale and affect a wider number of people. To undertake this research, patient conditions were derived from the pharmacy medical records of 15,812 patients. These conditions, as well as the patients attributes, were then run through a feature selection process to identify the important features for predicting diabetes. These variables were then used to train a random forest and artificial neural network achieving an accuracy of 77.1% and 76.4% respectively.

Keywords: Machine Learning · Diabetes · Pharmacy Medical Records.

1 Introduction

Diabetes is an epidemic disease affecting millions of people throughout the world. The World Health Organisation estimates that there are currently more than 240 million people living with the disease, and by 2030 it will be the seventh leading cause of death [9]. Diabetes can be split up into two main categories: type I and type II diabetes. Type I diabetes occurs when the body fails to produce enough insulin, and type II diabetes which occurs when the body becomes insulin resistant. Insulin is an important hormone produced by the pancreas, to send signals to the bodies cells to absorb glucose from the bloodstream. When cells fail to properly absorb glucose, glucose levels in the bloodstream rise. This can result in major health complications such as coronary artery disease, heart attack, stroke, nerve damage and blindness [15]. As such, early diagnosis of the disease is vital in order to better manage the disease and improve patient outcomes. In this study, type II diabetes will be specifically tackled.

Advances in machine learning have allowed researchers to develop a variety of predicative models for classifying diabetes [1]. The limitation with these models

is that they use clinical healthcare datasets, which are not easy to capture as they require patients to visit a clinical setting. This visit is generally done through the advise of a healthcare professional. This presents a two-fold problem: persons being tested have already been identified as at-risk and they might already have significant health consequences if they have been living with undiagnosed diabetes. Testing patients in a clinical setting also requires manual processing of test results. This is an expensive process as specialised equipment and healthcare professionals are needed in order to determine a diagnosis. Pharmacy medical records are a a passively generating data source, which are gathered whenever a patient visits their local pharmacy. This offers a readily available set of patient features, that can be used for ongoing diabetes screenings. This could greatly reduce the cost of preliminary screens and could lead to an earlier diagnosis.

In this study, a machine learning model for predicting diabetes is proposed using pharmacy patient medical records. For this work, the core research question is defined:

To what extent can pharmacy patient medical records be used to predict type II diabetes?

There are several challenges in answering this research question. These include: gathering a sufficient number of patient records, deriving suitable training features from the pharmacy dataset and devising a strategy in order to reduce complexity of the predictive model. In order to address these issues, and better answer the research question, the following objectives were devised: (i) Identify the most important patient features in the dataset for predicting diabetes, (ii) Construct a set of models using the features identified in objective one and (iii) Evaluate the models, and compare them against the current state of the art research.

In section two of this study the related works and state of the art models will be discussed. Section three will cover the methodology and implementation for the approach in this study and section four will evaluate the results of this study and discuss the implication of the results achieved.

2 Related Works

Advances in computer science have allowed for the crossover of medical expertise and machine learning for classifying diabetes. Models such as artificial neural networks, support vector machines, random forests, bayesian networks and logistic regression have been used to tackle this problem [1]. In the following section, random forests and artificial neural networks are identified as having the most accurate predictive ability for classifying diabetes. Neural networks work well when dealing with complex medical datasets where relationships between conditions are not always understood. Random forests offer similar advantages, as well as being suitable where training datasets are limited in sample size, which is often the case when dealing with clinical healthcare datasets.

2.1 Artificial Neural Networks

[2] have demonstrated the success of a backpropagation neural network in their approach for predicting diabetes. The paper proposes an 8-10-1 (1 hidden layer with 10 nodes) network, which employs a Levenberg-Marquardt algorithm. This method has an iterative process, which finds the local minimum of the model to best fit the data. Similarly, [12] apply a backpropagation approach for calculating the gradient of the cost function in a small-word feedforward neural network for classifying diabetes. Unlike the Lavenberg-Margquart algorithm demonstrated by [2], a bipolar sigmoid transfer function is used to activate the neurons. The sigmoid function optimises the model, making it a computational less intensive approach. In contrast, [12] configure their diabetes predictive neural network with 2 hidden layers, as opposed to 1, with 8 input nodes. In both instances, the researchers demonstrate how the proposed models achieve a high degree of accuracy of 93% and 81% respectively. As shown by [1], neural networks have outperformed bayesian networks for classifying diabetes patients. The researchers demonstrate that a naive bayes approach assumes independent between features, making it computationally less expensive than artificial neural networks. Although this reduces execution time, the researchers conclude that this assumption may lead to less accuracy, making the neural network a more favourable approach when dealing with predicting diabetes. A similar benchmarking exercise conducted by [10] also shows how an artificial neural network outperforms a logistic regression model for predicting diabetes. The authors configure a 3-layered perception neural network with 2 hidden layers and achieve an accuracy of 86% versus that of 78% achieved by the logistical regression model. [6] also apply a neural network to classify diabetes and use a genetically optimised algorithm for maximising the accuracy of the model. The genetic algorithm works to find the optimum weights and bias for the nodes by applying a fitness function to minimise the mean squared error between the predicted and actual classification of diabetes status during the training phase. This approach, coupled with feature selection results in a classification accuracy of 96%. A similar approach is taken in research conducted by [7], who use a convolutional neural network (CNN) for diabetes prediction. The CNN is a feedforward neural network, which has a multiple hidden layers, unlike the models configured by [2] and [10]. The authors illustrate the advantages of the CNN, which allows for meaningful insights into the relationship between different features, without any prior pre-processing. This omits the need for feature selection, such as that conducted by [7], thus reducing implementation time of the model.

2.2 Random Forest

Researchers have demonstrated the use of random forests in the prediction of diabetes with positive results when compared with other models. [11] illustrate this through their random forest model, which returned a better recall, precision and specificity when benchmarked against naive bayes and logistic regression. In this random forest model, an accuracy of 93% is achieved. Classification rates

in the study of [13] also show a high level of accuracy of 89%. Similarly to [11], the researchers show how random forests outperform logistic regression as well as support vector machines. The study also highlights the random forests ability to learn without any underlying assumptions of feature importance. In contrast to [13], [3] use an ensemble method combining classification and regression trees (CART) and random forest. CART is used to find the maximum average purity in splitting the nodes of the decision tree. The authors demonstrate how using this combined method overcomes accuracy problems encountered in other studies. Model optimisation was performed on 600 training datasets and 100 test datasets. As more trees are introduced the researches show how the classifier error rate decreases, with a maximum accuracy of 84% achieved. In contrast, [4] use a CART and random forest model in isolation and compare the results of the two models. The researchers found that the random forest outperforms the CART approach with an accuracy of 75% versus 65% respectively. Random forests have also been compared against adaptive boosting and iterative dichotmiser 3 algorithms. A study by [5] demonstrates this by benchmarking a random forest model for predicting diabetes against these algorithms. The researchers achieve a better prediction accuracy for the random forest across a number of different model configurations. The researchers conclude that because the random forest is more sensitive to early warning signs of diabetes, performance increases significantly as the volume of training data is increased when compared against the other models. The study finds a maximum classification accuracy of 84%, in contract to the next best model, the adaptive boosting model, which achieves an accuracy of 82%.

2.3 Related Works Findings

In the literature the use of pharmacy data has been under explored as a means of classifying diabetes. This gap in the research will be assessed in this study in order to determine how well pharmacy data can be used to tackle such problems. Many of the models in the literature also predefine the training features for diabetes classification. In contrast, this study implements a feature selection process, rather than eliminating certain variables from the beginning. This may uncover new knowledge about diabetes predictive features. The set of models used in this study use the best in class machine learning methods identified in the literature. A random forest and artificial neural network are applied. In this study, the random forest takes a similar approach to [3] and the artificial neural network is configured with a logistical activation function like [12], as this has proven to achieve the best classification rate for diabetes diagnosis. As with the [11] model, the number of trees chosen for the random forest is increased until no more significant accuracy is gained. For the artificial neural network a node pruning approach is also applied in order to find the optimum model configuration.

3 Methodology & Implementation

The approach taken in this study to address the research question was through the use of two machine learning models: a random forest and artificial neural network. In order to test the proposed dataset, patient medical records were obtained from 42 pharmacies across Ireland. In total 15,812 records were processed and transformed into the correct format for training the models. Once the models had been trained, a graphical user interface was created, where new data could be tested.

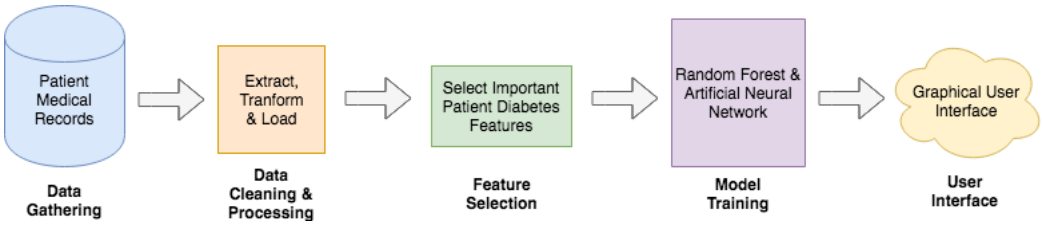

Fig. 1. Methodology pipeline from data gathering to finalised user interface

This research was conducted using the R open source programming language, through the RStudio integrated development environment [19]. The extract, transform, load process, model training and model application were all developed in R. The following sections will describe the implementation in detail and outline how the random forest and artificial neural network models were configured.

3.1 Feature Selection

In order to derive the most important features for the training models, the fscaret package was used [20]. Fscaret is designed for automatic feature selection through the use of a variety of models. By comparing the results of multiple models a smoothing affect is achieved for finding the important features of a classification problem. Each variable is scaled according to its mean-square error (MSE) and root-mean-square error (RMSE) for comparison. The models chosen for this feature selection implementation were: Gradient Boosting Machine (gbm), generalised linear model (glm), neural network (neuralnet), random forest (rf) and treebagging. Once the models were trained, the results were presented in a matrix listing the feature importance. Each variable was ranked based on its weighting across each of the models according to its MSE and RMSE. In order to decide which variables would be kept, and which ones would be ommited, a variable importance plot was produced and a cut-off point for the model features was defined.

3.2 Random Forest

The training dataset for the random forest consisted of 15,812 patient records. This was evenly split between diabetic and non-diabetic patients. The data was then partitioned into a 70/30 split for the training and testing phases. Using the randomForest package [21], the model was implemented and a series of decision trees were produced. By using this ensemble method, each tree produces a slightly different result. These results are then combined and the mode decision of all the trees is taken. By doing this, the issue of overfitting by a single decision tree is addressed. The variables used in the random forest training model were those that were defined during the feature selection stage. Using the bootstrap aggregation function, multiple random samplings of patient records were taken from the training dataset. In total, 30 bootstrap replications were taken to produce 30 individual decision trees. Beyond this point there was no significant performance gain from adding additional trees, only an increased computational cost. Once the model had been computed, the predict function was used on the testing dataset to classify each of the patient records as diabetic or non-diabetic. These result were then compared against the actual diabetes status for each patient. In order to better understand the model, the decision trees produced by the random forest were plotted using the partykit package [14].

3.3 Artificial Neural Network

The same training dataset was taken for the artificial neural network and was also partitioned into a 70/30 split for training and testing. The neuralnet package developed by [8] was used to implement the model. A backpropagation algorithm was used to retroactively recalculate the weights of each of the nodes. A variety of configurations were tested to optimise the number of hidden layers and the nodes within each layer. The models tested were all configured with 9 input nodes, as the number of training features selected remained the same for each iteration. Models with one hidden layer and two hidden layers were tested with no increase in accuracy being achieved by the two-layer configuration. For this reason, the one hidden layer model was selected to improve computational efficiency. The number of nodes within the hidden layer was also adjusted to find the optimal network performance. A series of tests and node pruning found that 4 nodes within the hidden layer achieved the highest accuracy results. This resulted in an 9-4-1 (one hidden layer with four nodes) neural network. The activation function, which is used to convert the input signal of each node into an output signal, was set to a logistic function for smoothing the results and the threshold for the error of the partial derivatives was set to 0.01, as this configuration achieved the greatest accuracy in the related works. The error of the neural network was calculated by the sum of squared errors. Once the model had been trained, its performance was measured using the test dataset. The model computed the diabetes status for each patient record, and the results were compared with the actual diabetes status for each patient to calculate the precision, recall, F1 score and accuracy of the model. As with the random forest, the neural network was plotted to visualise the model configuration.

4 Evaluation

In this section the outcome of the feature selection process is discussed, alongside an evaluation of the prevalence of these features in the current literature. The results of the random forest and artificial neural network are also presented and the models are tested against a number of performance metrics. These findings are then further explored and compared in the context of similar studies in the discussion section.

4.1 Feature Selection

In total, 84 features were available in the dataset during the feature selection process. The percentage for how much each feature contributed to the chosen models was calculated and summed across the models. The cut-off point was chosen at the first 9 features as the addition of subsequent features only had a minor contribution to the model. These feature selected were: Age, sex, diuretic medications (used to treat high blood pressure), intestinal secretion medications (used to help digest fats and regulate cholesterol), lipid-regulated drugs (used to treat cardiovascular diseases and alleviate high cholesterol), hypertension medications (used to treat high blood pressure), rectal disorder medication (used to treat rectal disorders), positive isotopic drugs (used to treat heart failure by increasing the strength of cardiovascular contractions) and laxatives (used to treat or prevent constipation).

4.2 Random Forest

The random forest was carried out using the variables identified in the feature selection process. In total, 30 decision trees were configured as part of the bootstrap aggregation process. The results of the random forest model are presented in Table 1 and Table 2. These show an overall accuracy for the model of 77.1%. The recall and the false positive rate were also plotted against each other to calculate the receiver operating characteristic curve (ROC). The results of the ROC curve show an area under the curve of .803 (Figure 2) for the random forest.

Table 1. Random forest model results

Model	**True Positive**	**True Negative**	**False Positive**	**False Negative**
Random Forest	1880	1779	492	593

Table 2. Random forest performance metrics

Model	**Precision**	**Recall**	**F1 Score**	**Accuracy**
Random Forest	.793	.760	.776	.771

4.3 Artificial Neural Network

The input features used in the neural network were also those identified during the feature selection process. A number of different parameters were tested for the neural network to find the optimal configuration. The number of hidden layers, and nodes within the hidden layer, were adjusted until the most accurate model of 1 hidden layer with 4 nodes was found. A 9-4-1 configuration for the model was defined.

The neural network is split up into three layers. The input layer represents where the features for the patients are fed into the model. Each feature is represented by an individual neuron in the input layer. The second layer is the hidden layer, which was configured with 4 neurons. The final layer is the output layer where the diabetes classification is predicted for the patients. The yellow neurons represent the values of the weighted connections between the neurons, which were initially randomised before converging at their final values. Each neuron in the input layer has a connection to the hidden layer with a corresponding weight. The sum of the value of each neuron and its connected neurons are added together and multiplied by their connection weights. This produces a bias value that is then put into an activation function, which transforms the value. The activation function then propagates through the network to produce a diabetes classification at the output layer. As the network is computed, the weights are iteratively adjusted through backpropagation to find the best fit for the model. The results of the neural network are illustrated in Table 3 and Table 4.

Table 3. Artificial neural network model results

Model	**True Positive**	**True Negative**	**False Positive**	**False Negative**
Neural Network	1785	1839	533	587

Table 4. Artificial neural network performance metrics

Model	**Precision**	**Recall**	**F1 Score**	**Accuracy**
Neural Network	.770	.753	.762	.764

Table 4 show a maximum classification accuracy of 76.4%. The ROC curve was also plotted for the neural network (Figure 2) and an area under the curve of .764 was calculated.

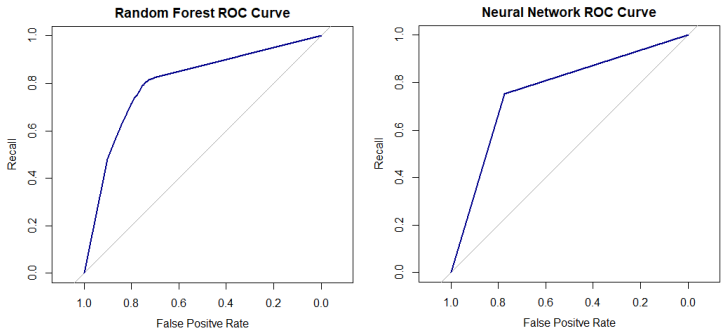

Fig. 2. Random Forest & Neural Network receiver operating characteristic curves

4.4 Discussion

The aim of this study was to see to what extent a pharmacys patient medical records could be used to aid the prediction of type II diabetes. By leveraging the capabilities of machine learning, the results have demonstrated a successful approach for a random forest and artificial neural network. The neural network achieved a classification accuracy of 76.4%, with the random forest achieving a slightly greater accuracy of 77.1%. These results present a new novel means of diabetes classification, not previously identified in the current literature. Although the accuracy of random forest and neural network did not outperform the state of the art models (Figure 5), these results still demonstrate how pharmacy data can give sufficient accuracy. However, the results obtained in this study are still around 16% less accurate than the best state of the art models. This is because clinical data has features such as family history and body mass index which are important indicators in such diseases.

Table 5. Comparison of accuracy of existing models against this study

Paper	**Method**	**Accuracy**
Alghamdi et al. (2017)	Random Forest	93%
Karatsiolis & Schizas (2012)		89%
Lpez et al. (2018)		84%
This Study		**77%**
Chung et al. (2014)	Neural Network	93%
Adavi et al. (2016)		86%
Erkaymaz et al. (2017)		81%
This Study		**76%**

The features identified in this study further validate the known diabetes risk factors identified in the related works chapter. The presence of patients taking lipid-regulated drugs and intestinal secretion drugs were identified as important features in the random forest and artificial neural network. These drugs are commonly used for treating high cholesterol, which is a major risk factor for diabetes. Similarly, hypertension and diuretic medications that are used to treat high blood pressure were also identified as a major indicator for diabetes by the models. This coincides with the known medical literature that persons with elevated blood pressure are at greater risk of developing diabetes. Age was also an important feature in both the random forest and artificial neural network in this study. Age is a commonly understood variable linked to diabetes risk, with an increase of age corresponding to an increased risk of developing the disease. The data used to train the models in this study showed a sharp increase in the number of patients with diabetes above aged 58, with a significantly lower number of instances of diabetes in younger patients.

Although the major variables identified in the models were consistent with the existing literature, an unexpected feature was identified. The presence of laxatives was not identified as a predictive feature for any of the other models reviewed in the related works chapter. However, studies by [17] and [18] have shown a link between diabetes and constipation. A consequence of diabetes is a high blood glucose level, which can ultimately lead to nerve damage of the digestive tract that can result in constipation. Laxatives are commonly used to treat constipation, thus potentially making them a useful feature for predicting diabetes as demonstrated in this study. Evidence from existing research, such as that by [16], also draws a relationship between constipation and high fat diets. These high fat diets can lead to high cholesterol, which further strengthens the finding that there may be a link between constipation and diabetes, as high cholesterol is an already known risk factor for the disease. The relationship between diabetes and constipation is relatively unexplored in the current literature and further research is needed in order to validate the prevalence of any such relationship.

This study suggests using the random forest and neural network models developed as a preliminary identification for diabetes patients, and not an outright diagnosis tool. Pharmacists should identify patients classified by the models as at-risk of diabetes and invite them in for a fasting blood glucose level test to confirm whether or not they do have diabetes, or if they fall within the pre-diabetes range. The results of these tests could then be introduced back into the training dataset to further improve the classification accuracy of the models and reduce the number of false positive diabetes classifications. Coupled with this data, the models could be refined even further to produce a three-way classification for identifying non-diabetic, pre-diabetic and diabetic patients.

5 Conclusion

The primary goal of this study was to answer the research question: "*To what extent can pharmacy patient medical records be used to predict to type II diabetes?*". This study found that a random forest and artificial neural network could be trained with pharmacy medical records to achieve an accuracy of 77.1% and 76.4% respectively. Although these results represent sufficient accuracy, they did not outperform the state of the art models, which use clinical healthcare datasets. Although there was a loss of accuracy of around 16% versus the state of the art models, this lack of detail is made up for by the wide availability of pharmacy medical records. This opens up a greater opportunity for the set of models in this study to be used on a wider audience outside of clinical healthcare settings. This model could be used as a preliminary diagnosis tool, and could identify patients who need further assessment. This could greatly reduce the cost and time taken to identify at-risk patients, as classifications could be run on a larger scale. As diabetes is a progressive disease, early diagnosis is critical for the long term wellbeing of patients. This model could help aid earlier diagnosis, which could ultimately lead to better patient outcomes.

5.1 Future Works

This study has demonstrated the potential of pharmacy data, and its application for predicting diabetes. Further research is needed in this area to see what other conditions can be predicted through machine learning methods using pharmacy medical records. It is also suggested that pharmacy data could be used in conjunction with clinical healthcare datasets to improve existing diabetes prediction models. More research is also needed to better understand the important risk factors of diabetic patients, and in particular the prevalence of laxatives and their relationship to diabetes, as identified in this study.

References

1. Alic, B., Gurbeta, L., Badnjevic, A.: Machine learning techniques for classification of diabetes and cardiovascular diseases. In: 2017 6th Mediterranean Conference on Embedded Computing. https://doi.org/10.1109/meco.2017.7977152
2. Joshi, S., Borse, M.: Title of a proceedings paper. In: Detection and Prediction of Diabetes Mellitus Using Back-Propagation Neural Network. https://doi.org/10.1109/icmete.2016.11
3. Sabariah, M. K., Hanifa, A., Sa, S.: Title of a proceedings paper. In: Early detection of type II Diabetes Mellitus with random forest and classification and regression tree (CART). https://doi.org/10.1109/icaicta.2014.7005947
4. Rallapalli, S., Suryakanthi, T.: Predicting the risk of diabetes in big data electronic health Records by using scalable random forest classification algorithm. In: 2016 International Conference on Advances in Computing and Communication Engineering (ICACCE). https://doi.org/10.1109/icacce.2016.8073762

5. Xu, W., Zhang, J., Zhang, Q., Wei, X.: Title of a proceedings paper. In: Risk prediction of type II diabetes based on random forest model. https://doi.org/10.1109/aeeicb.2017.7972337
6. Manajan, A., Kumar, S., Rohit, B.: Diagnosis of diabetes mellitus using PCA and genetically optimized neural network. In: 2017 International Conference on Computing, Communication and Automation (ICCCA). https://doi.org/10.1109/ccaa.2017.8229838
7. Zhang, J., Gong, J., Barnes, L.: HCNN: Heterogeneous Convolutional Neural Networks for Comorbid Risk Prediction with Electronic Health Records. In: 2017 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE). https://doi.org/10.1109/chase.2017.80
8. neuralnet: Training of Neural Networks, https://CRAN.R-project.org/package=neuralnet . Last accessed 16 June 2018
9. Global Report on Diabetes, https://bit.ly/1N8hn84. Last accessed 10 June 2018
10. Adavi, M., Salehi, M., Roudbari, M.: Artificial neural networks versus bivariate logistic regression in prediction diagnosis of patients with hypertension and diabetes. Medical Journal of The Islamic Republic of Iran **30**(312), (2016)
11. Alghamdi, M., Al-Mallah, M., Keteyian, S., Brawner, C., Ehrman, J., Sakr, S.: Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford Exercise Testing (FIT) project. PLOS ONE **12**(7), (2017)
12. Erkaymaz, O., Ozer, M., Perc, M.: Performance of small-world feedforward neural networks for the diagnosis of diabetes. Applied Mathematics and Computation **311**, 22–28 (2017)
13. Lopez, B., Torrent-Fontbona, F., Vinas, R., Manuel Fernandez-Real, J.: Single Nucleotide Polymorphism relevance learning with Random Forests for Type 2 diabetes risk prediction. Artificial Intelligence in Medicine **85**, 43–49 (2018)
14. Hothorn, T., Zeileis, A.: Partykit: A Modular Toolkit for Recursive Partytioning. Journal of Machine Learning Research **16**(1), 3905-3909 (2015)
15. Hanna, K., Guthrle, D.: Health-Compromising Behaviour and Diabetes Mismanagement Among Adolescents and Young Adults With Diabetes. The Diabetes Educator **7**(22), 223-230 (2001)
16. Taba Taba Vakili, S., Nezami, B., Shetty, A., Chetty, V., Srinivasan, S.: Association of high dietary saturated fat intake and uncontrolled diabetes with constipation: evidence from the National Health and Nutrition Examination Survey. Neurogastroenterology Motility **27**(10), 1389–1397 (2015)
17. Krishnan, B.: Gastrointestinal complications of diabetes mellitus. World Journal of Diabetes **4**(3), (2013)
18. Author, F.: Risk factors for chronic constipation based on a general practice sample. The American Journal of Gastroenterology **98**(5), 1107–1111 (2003)
19. A Language and Environment for Statistical Computing, http://www.R-project.org/. Last accessed 7 May 2018
20. Automated Feature Selection from 'caret', https://CRAN.R-project.org/package=fscaret. Last accessed 18 June 2018
21. Classification and Regression by randomForest, http://CRAN.R-project.org/doc/Rnews/. Last accessed 30 May 2018