EXPLORING DRIVERS FOR DIABETES IN USA: A DATA MINING APPROACH

Ganga Prasad Basyal[1]; David Zeng, PhD[2]; Omar-el Gayar,PhD[2]

[1]Dakota State University, [2]Dakota State University

Abstract

Diabetes Mellitus is one of the most prevalent chronic disease conditions in the United states. However, type-II diabetes mellitus (T2DM) is often encounters factors that can preemptively addressed through behavioral change. This research aims to 1) identify new factors associated with the onsets of the disease 2) Analyze the role of Data mining models in classifying the probability of having disease. Data mining models can provide the plausible support in the decision-making process by the physicians and can server as second opinion. The preliminary results show an encouraging accuracy rate of 91%.

Introduction

The choices and decision made by physicians in patient diagnosis are based on the knowledge and experience (Gandhi & Singh, 2015). Data Mining application on the other hand are the algorithms, models and workflow system which could be able to analyze much more data in less time as compared to humans. So as to minimize the time spent in analyzing the medical records and reports with less variations in interpretation, machine learning tools could be adopted. (Lee & Yoon, 2017).

Diabetes had taken the form of epidemic and is spreading at in increasing rate. To fight with the epidemic, we need to understand the trend and the prevalence of the disease and its effect to the citizens. The trends and prevalence show an increasing rate in diabetes in the last decade (Menke, Casagrande, Geiss, & Cowie, 2015). The rate of annual percentage change in United States for diabetes had seen and tremendous increase from 3.5% in 1980 to 8.3% in 2008 (Geiss et al., 2014). Although in the studies the role and importance of dietary intake evaluation is not evident (lee, Brancati, & Yeh, 2011).

Huge data is being produced in everyday basis in healthcare sector. Data comprises of medical records, patient visit, healthcare service providers claims and bills and personal health records. This information is not utilized to its full extent as the healthcare sector is abandoned with data but had less information interpretable information out of the data (Lee & Yoon, 2017). To extract the information or pattern from the data successful data mining techniques are required which will establish a logical and insightful information link from the database.

Feature Selection

Feature selection is one of the most important aspect of data mining. Features have directly proportional relation to Target variable. If positive of related variable is adopted it will affect positively and increase the model performance whereas, negative or non-relevant variable will penalize the model performance largely. Therefor we should be utmost careful while picking the right feature.

We had identified 10 feature those are highly relevant to the target and 9 are unique to our best knowledge. As the first aim of our research is to identify new feature for onset of diabetes we had achieved progress through discovering new feature. In the next section we will try to answer our second research aim i.e. the current model improvisation and new model adoption.

S.no	NHANES Code	Description	Target/Feature	Feature Identified (Is new)
1	BMXARMC	Arm Circumference (cm)	Feature	Yes
2	DRQSDIET	On Special Diet	Feature	Yes
3	RIDAGEMN	Age in months at the time of screening	Feature	No
4	CBD070	Money spent at supermarket/grocery store	Feature	Yes
5	CBD090	Money spent on nonfood items	Feature	Yes
6	CBD150	Time to get to grocery stores	Feature	Yes
7	DBD910	Number of frozen meals/pizzas in past 30 days	Feature	Yes
8	HSAQUEX	Source of Health Status Data	Feature	Yes
9	HUQ010	General health condition	Feature	Yes
10	RXDUSE	Taken prescription medicine, past month	Feature	Yes
11	DIQ010	Doctor told you have diabetes	Target	No

Table 1. Feature and Target Variable Selected.

Literature Review

We analyzed three closest studies to the best of our knowledge, those used any data mining techniques for diabetes prediction. The first research significantly utilize Support Vector Machine (SVM) using 10-fold cross validation. SVM is one of the most widely used traditional classification model. The study was focused more on building the model which was a web-based tool. The AUC acquired by implementing the model was 83.47 on test set for classification scheme -1 and 73.18 on classification scheme- II with test dataset.

Second Study focus on identification of onset for type 2 diabetes with respect to middle-aged subject with metabolic syndrome(Ozery-Flato et al., 2013). This research is more focused and narrow in terms of population study. In the research author implemented type 2 diabetes metabolic (T2DM) model using logistic regression model.

Third research focused on the ensemble classifier for predicting type-2 diabetes (Semerdjian & Frank, 2017). Author in the research used the NHANES data set. For feature selection they referred to study already done for feature selection and took the same variable. The study they referred was (Yu et al., 2010). In this research they introduced the ensemble classifier comprising of 5 different classification model

After analyzing all three researches, it is explicit that first research limits in model adoption. Having multiple models could helped in validation of results and contribute in deep analyzation, whereas second study limit at the number of observations used and the results obtained was generated through one model which again raise the question of model validity. Third study comprises of widely used classification model but the feature selection was not evident. Therefor, we need a Model encompassing the limitation from the previous researches and adopt combination of new model along with relevant feature selection methods. Selecting the features those are relevant to the target will highly boost the performance and vice-versa. Incorporating this would result in model which will have higher impact and enhanced accuracy in prediction.

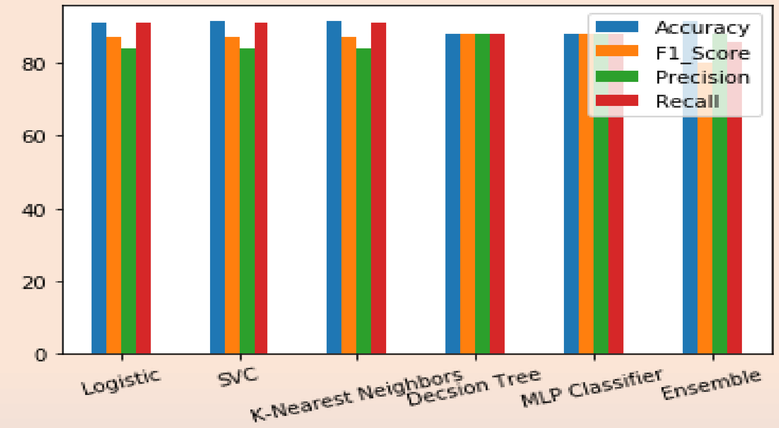
Reference

Babu, S., Vivek, E. M., Famina, K. P., Fida, K., Aswathi, P., Shanid, M., & Hena, M. (2017). Heart disease diagnosis using data mining technique. In *2017 International conference of Electronics, Communication and Aerospace Technology (ICECA)* (pp. 750–753). Coimbatore: IEEE. https://doi.org/10.1109/ICECA.2017.8203643

Brown, G., & White, E. (2017). An Investigation of Nonparametric Data Mining Techniques for Acquisition Cost Estimating. *Defense Acquisition Research Journal*, *24*(2), 302–332. https://doi.org/10.22594/dau.16-756.24.02

Preliminary Results and Analysis

Support Vector Classifier had achieved highest recall level of 91% and 91.41% accuracy. Having high recall value indicate the high generalization validity new dataset. Whereas, Ensemble model along with SVC and K-Nearest Neighbor had achieved the highest accuracy.



Model/ Accuracy	Precision	Recall	F1- score	support	Accuracy
Logistic Regression	0.84	0.91	0.87	1899	91.25%
Support Vector Regressor	0.84	0.91	0.87	1899	91.41%
K-nearest neighbors	0.84	0.91	0.87	1899	91.41%
Decision Tree	0.88	0.88	0.88	1899	87.94%
MLP classifier	0.88	0.88	0.88	1899	87.94%
Ensemble Average (LR/DT/SVR/k-NN)	0.88	0.86	0.80	1899	91.41%

Table 2. Precision Accuracy and Recall Matrix

Conclusions & Future Research

In our research we identified 9 completely unique set of variables. These features are selected based on feature selection technique called Recursive feature elimination (RFE) providing the ranks based on their correlation with the target variable. We implemented the new features into the previously used models along with Artificial Neural Network and achieved higher accuracy level. Our ensemble model produced the highest percent of accuracy among all individual models.

We implemented one feature selection method and would like to adopt more methods to observe the difference in behavior. Our Neural network model has provided accuracy of 88 % which is average. As we had implemented a very basic and simple form of neural network in future research we can retrain the neural network with adding more layers to it and making it deeper. The more we add the layer, deeper and complex the model become. We will use the hyperparameter for fine tuning the model through Sigmoid or SoftMax functions