Predicting Prediabetes Using Simple a Multi-Layer Perceptron Neural Network Model

Oana VÎRGOLICIa,1 and Horia VÎRGOLICIb

aAcademy of Economic Studies, Bucharest, Romania
b"Carol Davila" University of Medicine and Pharmacy, Bucharest, Romania

Abstract. In over 60% of patients with prediabetes, the evolution to diabetes can be stopped by changing lifestyle. Application of prediabetes criteria existing in accredited guidelines is very useful, representing an effective way to avoid prediabetes and diabetes. Although these guidelines imposed by the international diabetes federation are constantly updated, many doctors do not apply, mainly due to lack of time, the recommended steps for diagnosis and treatment. In this paper, a multi-layer perceptron neural network model for prediabetes prediction is proposed, based on a dataset with 125 persons (men and women), with the following features: gender (S), serum glucose (G), serum triglycerides (TG), serum high-density lipoprotein cholesterol (HDL), waist circumference (WC) and systolic blood pressure (SBP). The output feature in the dataset (prediabetes or not) was based on a standardized medical criterion named Adult Treatment Panel III Guidelines (ATP III), which specifies that prediabetes diagnostic can be establish if at least three of five parameters are outside the scale of their normal values. Satisfactory results were obtained in evaluating the model.

Keywords. prediabetes, deep learning, multi-layer perceptron, neural network model

1. Introduction

Researchers emphasize that in over 60% of patients with prediabetes, the evolution to diabetes can be stopped by changing lifestyle [1,2].

The prevalence of diabetes mellitus was 11.6% and of prediabetes was 16.5% in the Romanian population aged 20-79 years, according to PREDATORR (PREvallence of DiAbetes mellitus, prediabetes, overweight, Obesity, dyslipidemia, hyperuricemia and chronic kidney disease in Romania) study [3]. The highest percentage of prediabetes was in women, in the 60-79 year age group. Obesity, abdominal obesity, dyslipidemia, low education level, and a family history of diabetes were correlated with dysglycaemia.

In this paper a simple deep learning approach for prediabetes prediction is proposed, implemented in Python. The dataset was obtained from patients hospitalized
in several clinics in Bucharest, Romania. Dataset contains values of some usual parameters. The diagnostic of prediabetes is based on Adult Treatment Panel (ATP III) medical guidelines [4]. The proposed model, for different values of parameters, was evaluated in terms of accuracy.

2. A brief literature review

A brief review presents some of the models which use neural networks (but not only) or features of the dataset which are similar to those used in our study.

Khanam and Foo used NB (Naive Bayes), SVM (Support Vector Machine), Linear Regression, Adaboost, RF (Random Forest), kNN (k-Nearest Neighbors), DT (Decision Tree) and NN (Neural Network) with different hidden layer and compared their results with other results. The attributes that are used are Pregnancy, BMI, Insulin level, Age, Blood pressure, Skin thickness, Glucose, Diabetes pedigree function, and Outcome (0 means non-diabetes, and 1 implies diabetes). They used Weka, and data mining software tool for the diabetes dataset’s performance analysis [5].

Dewangan and Agrawal proposed a system for the diagnosis of diabetes using Bayesian classification and multilayer perceptron [6]. Data was classified into diabetic and non-diabetic. PIDD (Pima Indian Diabetes Dataset) was used. Accuracy of 81.89% was achieved.

Perveem et al. used a dataset obtained from the Canadian Primary Care Sentinel Surveillance Network, which includes: systolic blood pressure (SBP), diastolic blood pressure (DBP), HDL, triglycerides (TG), BMI (Body Mass Index), fasting blood sugar (FBS), and gender [7]. They used Bootstrap aggregating, Adaptive Boosting, and DT model. They found for better accuracy, Adaboost can be applied to predict diseases like diabetes, coronary heart disease, and hypertension.

3. Methodology

Dataset contains 125 patients (men and women), with the following features: gender (S), serum glucose (G), serum triglycerides (TG), serum high-density lipoprotein cholesterol (HDL), waist circumference (WC) and systolic blood pressure (SBP). The ethical norms regarding the confidentiality of patient data were respected. There were no missing values in the dataset.

Each patient is classified with prediabetes or not (feature named Output, whom values are 1 for prediabetes – 90 cases and 0 for no prediabetes - 35 cases), based on ATP III criterion, which establishes that at least three values from five should be out of normal range (TG ≥ 150 mg/dl, WC ≥ 102/88 cm for men/women, G ≥ 110 mg/dl, HDL < 40/50 mg/dl for men/women and SDP ≥ 130 mmHg) [4]. All these values are specific for population in USA and Europe (Europids).

A simple multi-layer perceptron neural network approach is proposed, written in Python and based on Keras. The following modules were imported: Sequential from tensorflow.keras.models, Dense from tensorflow.keras.layers, train_test_split from sklearn.model_selection and StandardScaler from sklearn.preprocessing.

Dataset was split in 80% for training and 20% for testing. A sequential model, using different values for parameters, was built. The model is based on scaled data
(made by StandardScaler() function) and non-scaled data. A comparison between all these versions, in terms of accuracy, is made.

4. Results and discussion

The sequential model having one, two and three hidden layers respectively, was built. Activation function was chosen ‘relu’ for all hidden layers and ‘sigmoid’ for the output layer.

The optimizer was chosen, in turn: Adam, Adadelta, Adagrad, Adamax, Nadam, RMSprop, FTRL and SGD. The loss function was chosen ‘binary_crossentropy’, which the most suitable for binary classification.

There were no significant differences in performance for the model with one, two or three hidden layers.

In Table 1 is shown accuracy for different values of optimizer and loss function, for model using scaled data. The model trained and tested on unscaled data has accuracy below the accuracy for model trained and tested on scaled data and those values are not represented in the table.

Table 1. Comparison between models having different optimizer and loss function, applied to scaled data, from the best to the worst accuracy

<table>
<thead>
<tr>
<th>Optimizer</th>
<th>Loss</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adam</td>
<td>binary_crossentropy</td>
<td>96%</td>
</tr>
<tr>
<td>RMSprop</td>
<td>binary_crossentropy</td>
<td>96%</td>
</tr>
<tr>
<td>Adamax</td>
<td>binary_crossentropy</td>
<td>88%</td>
</tr>
<tr>
<td>Nadam</td>
<td>binary_crossentropy</td>
<td>84%</td>
</tr>
<tr>
<td>Adagrad</td>
<td>binary_crossentropy</td>
<td>80%</td>
</tr>
<tr>
<td>SGD</td>
<td>binary_crossentropy</td>
<td>80%</td>
</tr>
<tr>
<td>Adadelta</td>
<td>binary_crossentropy</td>
<td>72%</td>
</tr>
<tr>
<td>FTRL</td>
<td>binary_crossentropy</td>
<td>68%</td>
</tr>
</tbody>
</table>

The loss and accuracy for train and test data are shown in Figure 1, for the best model, and in Figure 2, for the worst model.

![Figure 1. Accuracy and loss for model using ‘adam’ optimizer](image-url)
5. Conclusion

Early detection of prediabetes is the essential step in preventing diabetes. Doctors, which are pressed by time, can find in predicting models very useful tools. The features in dataset are based on systolic blood pressure and waist circumference and on usual blood tests, which are cheap and easily to be measured (serum glucose, serum triglycerides, serum HDL-cholesterol).

The best model presented in the paper has an accuracy of 96%, which is very encouraging, but we are aware that principal limitation of the paper is related to the relative small number of records in dataset.

References