LITREATURE SURVEY

By

M.Abimanyu

N.Gowthaman

C.S.KrishnaPrakash

B.SatheeshKumar

Prediction of Diabetes using Machine Learning

Abstract

Over 30 million people in India are suffering from diabetes and many others are under the risk. Thus, early diagnosis and treatment is required to prevent diabetes and its associated health problems. This study aims to assess the risk of diabetes among individuals based on their lifestyle and family background. The risk of diabetes was predicted using different machine learning algorithms as these algorithms are highly accurate which is very much required in the health profession. Once the model will be trained with good accuracy, then individuals can self-assess the risk of diabetes. In order to conduct the experiment, 952 instances have been collected through an online and offline questionnaire including 18 questions related to health, lifestyle and family background. The same algorithms were also applied to the Pima Indian Diabetes database. The performance of Random Forest Classifier is found to be most accurate for both datasets.

1. Introduction

Diabetes, also known as diabetes mellitus (DM), is a set of metabolic issues identified by high blood glucose levels over a prolonged duration of time. Symptoms of high glucose incorporate excessive urination, always feeling thirsty and increased hunger [1]. If not treated on time, diabetes can cause serious health issues in an individual such as diabetic ketoacidosis, hyperosmolar hyperglycemic state, or even lead to death. This may lead to lifetime

- Diabetes Mellitus Type-1 is characterized by pancreas generating insulin less than what is required by the body, a condition also called "insulin-subordinate diabetes mellitus" (IDDM). People suffering from type-1 DM require external insulin dosage to make up for the less insulin produced by the pancreas.
- Diabetes Mellitus Type-2 is marked by the body resisting insulin as the body cells react differently to insulin than they normal would. This may ultimately lead to no insulin in the body. This is otherwise called "non-insulin subordinate diabetes mellitus" (NIDDM) or "adult starting diabetes". This type of diabetes is commonly found in people with high BMI or those who lead an inactive lifestyle.
- Gestational diabetes is the third principle structure that is observed during pregnancy.

Generally, for a normal human being, glucose levels range from 70 to 99 milligrams per deciliter. A person is considered diabetic only if the fasting glucose level is found to be more than 126 mg/dL. In the medical practice, a person having a glucose concentration of 100 to 125 mg/dL is considered as pre-diabetic [4]. Such a person is prone to the development of type 2 diabetes. Over the years, it has been found that people with the following health characteristics face a greater risk against diabetes:

- A Body Mass Index value greater than 25
- Members of the family suffering from diabetes
- People with HDL cholesterol concentration in the body less than 40 mg/dL
- Prolonged hypertension having gestational diabetes
- People who have suffered from polycystic ovary disorder in the past
- People belonging to ethnic groups like African American, or Native American, or Latin American, or Asianpacific aged over 45 years
- Having an inactive lifestyle

When a doctor diagnoses that an individual has prediabetes, they suggest the individual better their lifestyle.

Adopting a fitness regime and a good diet plan can help prevent diabetes [5].

This research aims to determine the risk of development of diabetes in an individual. The subsequent sections of the article consist of related work in section 2. The brief description of machine learning algorithms applied is presented in section 3. The methodology is described in section 4 while results in section 5. Section 6 summarizes the conclusion.

2. Related work

In India, diabetes is a widespread problem as more than 70% of the adult population is suffering from this disease. Various researchers have worked to predict symptoms of diabetes by applying different approaches such as machine learning and data mining [11]. Few of them have also applied neural network and genetic algorithm. Since the problem of prediction of diabetes is supervised in nature, the supervised methods of machine learning, data mining and ANN have been applied by many.

Some closely related works are discussed in this section. Many of the research studies have used Pima Indians Diabetes Dataset (PIDD) for diabetes prediction. Machine learning methods and Weka tool were applied by [13, 14, 16, 17, 20, 21, 23]. The different approaches applied by researchers can be broadly classified as machine learning methods, data mining techniques, hybrid methods and neural network or genetic algorithms.

Swapna et. al. in [12] used deep learning methods on electrocardiogram (ECG) signals for detection of diabetes. Specifically, convolution neural network and long short-term memory has been used by them and then features were extracted by support vector machine. As a result, they found a very high accuracy of 95.7%.

Sisodia et. al. in [13] applied three machine learning methods i.e. decision tree (DT), naïve based (NB) and support vector machine (SVM) on PIDD in order to predict the diabetes. Naïve bayes classifier was found to be 76.30% accurate.

Han Wu et. al. in [14] applied data mining techniques (i.e. improved kNN and logistic regression) to accurately predict up to 95.42% the risk to an individual of developing type 2 diabetes. The modification was done by selecting value of initial seed point experimentally. The initial seed point was selected by conducting 100 experiments in which they selected smallest value of 'within cluster sum of squared errors.'

Meng et. al. in[15] compared logistic regression, artificial neural network (ANN) and decision tree (DT) for identifying the risk of diabetes and prediabetes based on 12 risk factors which included education level, work stress, BMI, age, sleep duration, gender, marital status, family history of diabetes, coffee drinking, preference to salty foods, physical activity, and consumption of fish. DT was found to provide best results among the three methods.

Choubey et. al. in [16] applied a hybrid algorithm using genetic algorithm (GA) and radial basis function neural network (RBFNN), wherein first GA is applied for feature selection then RBFNN is applied for classification. Their findings were that hybrid method performed better than RBFNN alone.

Tigga et. al. in [17] applied logistic regression on PIDD for diabetic prediction and found number of pregnancies, BMI and glucose level are most significant variables for diabetes prediction among all features in PIDD.

Huang et. al. in [18] did feature selection and classification of diabetes by applying naïve bayes, IB1 and C4.5 algorithms. The study concluded that patient age, diagnosis duration, need of insulin and diet control are most important factors for blood sugar control. Some other factors are also affecting results that are type of care, home monitoring and importance of smoking.

Saravana et. al. in [19] collected raw data from various places in form of Electronic Reports (EHR) that may be clinical reports, prescriptions given by doctors, diagnostic centre reports, pharmacy related information, and data asked by insurance personals. All this information collectively put in a map reduce to exact features which are directly related to diabetes.

Nongyao et. al. in [20] compared four classification techniques i.e. decision tree, ANN, logistic regression and naïve bayes. Further bagging and boosting were applied on all and random forest was also included. The maximum accuracy achieved by all was in between 84% and 86%.

Zou et. al. in [21] applied Random Forest, J48, ANN for classification after the feature reduction is done by unsupervised methods: Principle Component Analysis (PCA) and Minimum Redundancy Maximum Relevance (mRMR) methods. Accuracy for mRMR is found to be better than PCA with all features.

Perveen et. al. in [22] were concerned in finding risk of metabolic syndrome and diabetes. For prediction Naïve Bayes and J48 (C4.5) decision tree model were applied and the balancing of training set was done by k-medoids sampling. In their study, NB outperformed the others.

Rahman et. al. in [23] summarises the effect of various data mining techniques for diabetes diagnosis. For the prediction purpose, Multilayer Perceptron (MLP), Bayes Classification, J48graft, JRip (RIP- PER), Fuzzy Lattice Reasoning (FLR) classification methods were applied. J48graft was found most accurate.

Choi et. al. in [24] applied machine learning algorithms on patients having history of non-diabetes having cardiovascular risk. Five years data has been collected in form of EMR from Korea University Guro Hospital. Then, machine learning methods were applied with 10-fold cross validation. Highest accuracy was obtained in logistic regression model.

3. Brief description of Machine Learning Classification Techniques

3.1. Logistic Regression Method

Logistic regression is a sort of supervised learning which estimates the connection between a binary dependent variable and at least one independent variable by evaluating probabilities with the help of sigmoid function. In contrary to its name, logistic regression is not used for regression problems rather is a type of machine learning classification problem where the dependent variable is dichotomous (0/1, -1/1, true/false) and independent variable can binominal, ordinal, interval or ratio-level. The sigmoid/logistic function is given as [6]:

$$y = \frac{1}{1 + e^{-(1)x}}$$

Where, y is the output which is the result of weighted sum of input variables x. If the output is more than 0.5, the output is 1 else the output is 0.

3.2. K- Nearest Neighbor Classifier

K-Nearest Neighbor (KNN) method can be used to solve problems pertaining to both regression as well as classification, though it is generally being used to solve classification problems in business. Its major advantage is simplicity of translation and low computation time. In figure 1, the points (2.5, 7) and (5.5, 4.5) will be allocated to any one of the clusters. The KNN uses Euclidean distance function to find distances between existing data points and any new data point. Thus, (2.5, 7) will belong to the green cluster, whereas, (5.5, 4.5) will belong to the red cluster [7].

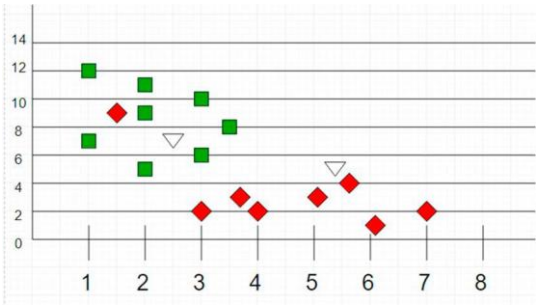

Fig. 1. KNN example [7]

3. 3. Support Vector Machine (SVM)

SVM is a supervised classifier in machine learning algorithms that can be used both for regression and classification. It is majorly applied in solving classification problems. The goal of SVM is to classify data points by an appropriate hyperplane in a multidimensional space. A hyperplane is decision boundary to classify data points. The

hyperplane classifies the data points with maximum margin between the classes and the hyperplane. Figure 2 shows support vector machine classification [8].

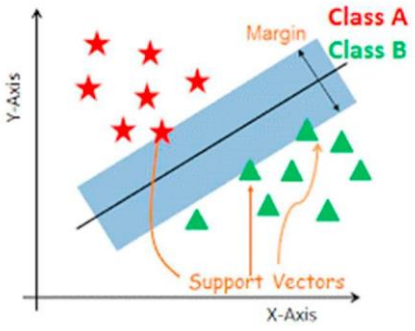

Fig. 2. Support Vector Machine [8]

3. 4. Naïve Bayes Classification Method

Naïve bayes classification method is a probabilistic machine learning algorithm based on Bayes theorem described in probability. Even with its simplicity it outperforms other classifiers; hence, it is one of the best classifiers. The Bayes theorem for calculating posterior probability is given below [9]:

$$p(c|x) = \frac{p(x|c)\,p(c)}{p(x)} \quad (2)$$

Where,
P(c|x) = Posterior Probability
P(x|c) = Likelihood
P(c) = Class Prior Probability
P(x) = Predictor Prior Probability

3. 5. Decision Tree Classification Method

A Decision Tree is works on the principle of decision making. It can be described in form of tree and provides high accuracy and stability. Figure 3 illustrates a decision tree [10].

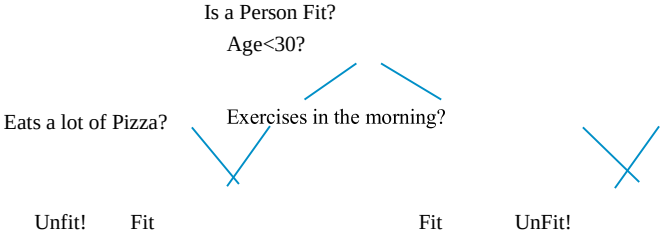

Fig. 3. Decision Tree Example

3.6 Random Forest Classification

The Random forest classifier creates multiple decision tree from randomly selected subset of training dataset shown in figure 4. Then it aggregates the votes from different decision trees to decide the final class of test objects [29].

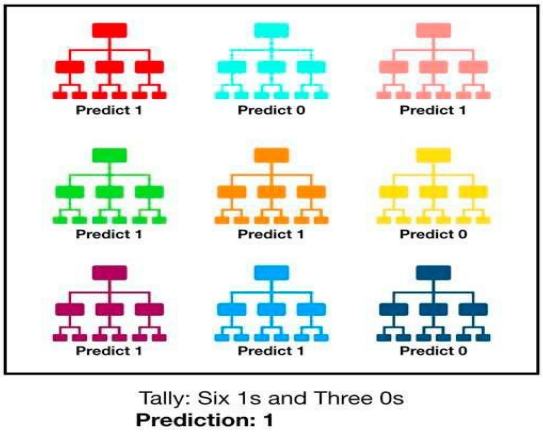

Fig.4. Random Forest Classification

4. Methods

4.1. Data Description

In this study, total 952 participants are selected aged 18 and above, out of which 580 are males and 372 are females. The participants were asked to answer a questionnaire shown in Table1 which was self-prepared based on the constraints that could lead to diabetes. In order to verify the validity of model same experiments were performed on another database called PIMA Indian Diabetes database [28] shown in Table1. Figure 6 shows sample dataset collected through questionnaire.

Table 1. Questions and possible answer in Questionnaire.

	Our Dataset		Pima Dataset	
	Parameters	Instances	Parameters	Instances
	Total participants	952	Total participants	768
1	Age	18 or above	Age	21 or above
2	Gender: • Male • Female	580 372	Gender: □ All female	768
3	Family history with diabetes	Yes/ No	Pregnancies	Numeric
4	Diagnosed with high blood pressure	Yes/ No	Glucose	Plasma glucose concentration a 2 hours in an oral glucose tolerance test.
5	Walk/run/physically active	• None • Less than half an hour • More than half an hour • One hour or more	Blood pressure	Diastolic blood pressure (mm Hg)
6	BMI	Numeric	Skin thickness	Triceps skin fold thickness (mm)

7	Smoking	Yes/No	Insulin	2-Hour serum insulin (mu U/ml)
8	Alcohol consumption	Yes/No	BMI	Body mass index (weight in kg/(height in m))
9	Hours of sleep	Numeric	Diabetes pedigree function	Diabetes pedigree Function
10	Hours of sound sleep	Numeric	Outcome	• Diabetic – 268 • Non- diabetic - 500
11	Regular intake of medicine?	Yes/No		
12	Junk food consumption	Yes/No		
13	Stress	□ Not at all • Sometimes • Often • Always		
14	Blood pressure level	High/normal/low		
15	Number of pregnancies	Numeric		
16	Gestation diabetes	Yes/No		
17	Frequency of urination	□ Not much • Quite much		
18	Diabetic?	□ Diabetic - 267	□ Non diabetic - 685	

	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Age	Gender	Family_Di	highBP	Physically	BMI	Smoking	Alcohol	Sleep	SoundSlee	RegularMed	JunkFood	Stress	BPLevel	Pregancie	Pdiabetes	UriationFr	Diabetic
2	50-59	Male	no	yes	one hr or	39	no	no	8	6	no	occasiona	sometime	high	0	0	not much	no
3	50-59	Male	no	yes	less than	28	no	no	8	6	yes	very ofter	sometime	normal	0	0	not much	no
4	40-49	Male	no	no	one hr or	24	no	no	6	6	no	occasiona	sometime	normal	0	0	not much	no
5	50-59	Male	no	no	one hr or	23	no	no	8	6	no	occasiona	sometime	normal	0	0	not much	no
6	40-49	Male	no	no	less than	27	no	no	8	8	no	occasiona	sometime	normal	0	0	not much	no
7	40-49	Male	no	yes	none	21	no	yes	10	10	no	occasiona	sometime	high	0	0	not much	yes

Fig. 5. Dataset used in this study.

4.2. Design and Implementation

For the purpose of the study, RStudio was used for implementation and R programming language was used for coding. Machine learning algorithms like logistic regression, k-nearest neighbour, support vector machine, naive bayes classifier, decision tree and random forest classifications were implemented on the dataset collected and the

Pima dataset in order to predict diabetes. All these predictions from each classifier are then compared with each other. The following are the steps to apply machine learning algorithm, shown in figure 5.

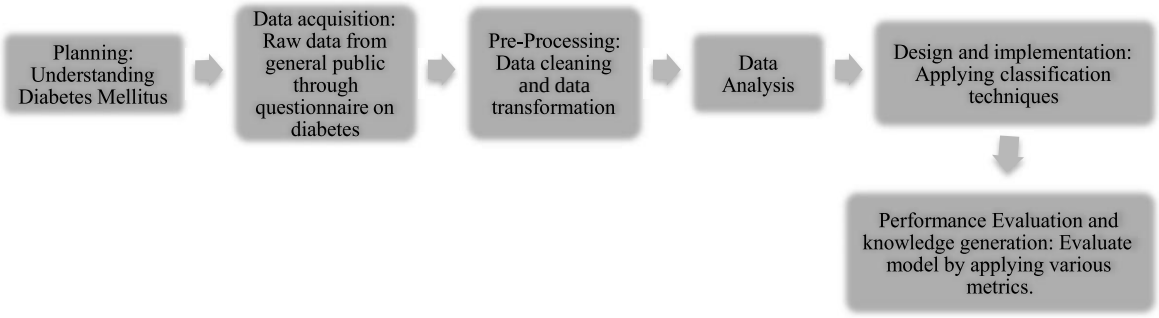

Fig. 6 . Flowchart of Research.

6. Conclusion

One of the global health issues is to identify the risk of diabetes at its early phase. This study attempts to structure a framework which forecasts the risk pertaining to diabetes mellitus type 2. In this paper, six machine learning classification methods were implemented, and their results were compared with different statistical measures. Tests were performed on the dataset collected through online and offline questionnaires consisting 18 questions relevant to diabetes. Also, same algorithms were applied on PIMA database. The experimental result shows that the accuracy of Random Forest of our dataset is 94.10% which is the highest among the rest. Random forest is also giving highest accuracy for PIMA dataset. Among six different machines learning algorithms applied, all the models produced good results for some parameter like precision, recall sensitivity etc. The observation made by figure 7 that ‘Age’, ‘Family diabetes’, ‘Physically active’, ‘Regular Medicine’ and ‘Pdiabetes’ or gestation diabetes has highest significance among all variables. These parameters have greater impact on predicting diabetes than the rest.

This result can be used in future to predict any other ailment. This study still holds a scope for further research and improvement including other machine learning algorithms to predict diabetes or any other disease.