Design and Development of Diabetes Prediction Model

[1]Chandan Kumar, [2]Nanhay Singh
[1]Computer Science & engineering,
[1]Ambedkar Institute of Advanced Communication technologies & Research, Geeta colony, Delhi, India

Abstract: Nowadays, diabetes has become a common diseases to the mankind from young to old persons. The growth of the diabetic patients is increasing day by day due to various causes such as toxic and chemical contents mix with the food, bacterial or viral infections, bad diet, change in life style, environmental pollution etc. The population with diabetes in 2016 was estimated to be 415 million worldwide. This figure is predicted to rise to 642 million by 2040. The data analytics is a process of examining and identifying the hidden patterns from large amount of data to give conclusions. In healthcare, this analytical method is done using machine learning algorithms for analyzing medical statistics to build the system getting to know models to perform scientific diagnoses. This paper presents a diabetes prediction model to diagnosis diabetes. This paper explores the approaches to improve the accuracy in diabetes prediction using medical data with various machine learning algorithms and methods such as Neural network, support vector machine, logistic regression, KNN etc.

Index Terms **- Diabetes, Machine learning, Neural Networks, hyper-plane, blood sugar, support vector machine.**

NOMENCLATURE

IGT (Impaired Glucose Tolerance) – It is a condition in which blood sugar is high, but not as high as to be type 2 diabetes.

IFG (Impaired Fasting Glucose) – It is a condition of pre-diabetes, in which a person's blood sugar level during fasting are consistently above the normal range.

OGTT (Oral Glucose Tolerance Test) – It is a method which ca help to diagnose instances of diabetes mellitus.

I. INTRODUCTION

Diabetes is a fast developing disorder most of the people even among the teen in the complete international. Diabetes is a common diseases found in all over the world due to stress, food and abnormal routine. Diabetes is due to growth level of the sugar within the blood (High blood glucose) [1]. Diabetes mellitus, usually called diabetes, is a set of metabolic disorders wherein there are high blood sugar level over a prolonged period [2]. Diabetes leads numerous illnesses which includes cardiovascular headaches, renal issues, retinopathy and foot ulcers and so on.

1.1. Types of Diabetes

Type 1 diabetes- It is a chronic condition in which pancreas produces a little or no insulin. It is known as insulin dependent diabetes [3].

table I

Type 1 diabetes	
Target level	4-8 mmol/L before meals <10 mmol/L two hour after starting meals

Type 2 diabetes – In it body blood sugar processes ability affected. It begins with insulin resistance, a condition in which cells fail to respond to insulin properly [3].

table II

Test	Normal	IFG or IGT	Type 2 diabetes
Hemoglobin A1c level, %	<5.7	5.7-6.4	>6.5
Fasting plasma glucose level mmol/L mg/dl	 <5.6 <100	 5.6 100-125	 >7.0 >126
OGTT result mmol/L mg/dl	 7.8 <140	 7.8 <140	 >11.1 >200

Gestational diabetes – It is a form of high blood sugar that affects pregnant women [3].

table III

Fasting test	
Normal sugar level	<6.1 mmol/L – 110(mg/dl)
impaired fasting glucose	>6.1 mmol/L – 110 (mg/dl)
Impaired glucose tolerance	<7.0 mmol/L – 126 (mg/dl)
Diabetes mellitus	>7.0 mmol/L – 126 (mg/dl)

II. LITERATURE REVIEW

This section opinions numerous studies works which can be associated with the proposed paintings. Mohammed Abdul Khaleel et al performed a survey on machine getting to know techniques on medicinal data for finding locally recurrent sicknesses. The main aim of this survey is to analyse the machine learning techniques required for medicinal data analysis that is especially used to discover locally recurrent diseases such as heart lung cancer, breast cancer [4]. Kaveeshwar, S.A., and Cornwall, J., conducted a survey on current state of diabetes in India and how it is affected to them [1]. Chunhui Zhao et al presented a system for Subcutaneous Glucose Concentration prediction. This proposed model can predict Type 1 diabetes mellitus [5]. K. Srinivas et al developed applications of machine learning techniques in health care and prediction of heart attacks. This research used medicinal profiles such as age, intercourse, blood strain and blood sugar and predicted the chance of patients getting a coronary heart and kidney troubles [6]. M. Durairaj and V. Ranjani discussed the ability use of classification based totally machine studying strategies inclusive of rule based strategies ,decision tree algorithm, naïve bayes and artificial neural community to the huge volume of healthcare records [7].

III. PROPOSE WORK

I. To create a standard model for diabetes prediction.
II. To enhance model accuracy using boosting.
III. To enhance model stability using bagging.
IV. To create a model for very large data set.

IV. TECHNICAL APPROCHES OF PREDICTIVE MODELING

Technical analytics encompasses a selection of statistical techniques from predictive modelling, machine mastering and statistics mining that examine modern and historical facts to make prediction approximately future or otherwise unknown occasions.
There are many techniques that may be used to create a model.
The most common ones are –

- Artificial Neural Network
- Support Vector Machine
- Decision tree
- K-nearest neighbor
- Logistic Regression
- Gradient boosting

V. METHODOLOGY

We needs numerous kinds of data to generate predictive fashions. Data is the fuel that drives the analytics system. There are styles of statistics are required within the version development.

1. Predictor records – It is also called predictor variables. This sort of records we need to make predictions i.e. Statistics that could feature inside the predictive model. For example – People profits, age, growth fee and so forth.
2. Behavioral information - It is likewise referred as outcome facts. It is the behavior that we want to are expecting.

In order to build a predictive version, the development pattern needs to contain both kinds of facts. An suitable mathematical/ statistical technique is then applied to decide what courting exist between the predictor facts and conduct records. The relationships which might be located are contained within the ensuing version.

VI. DIABETES PREDICTION USING MEDICAL DATA

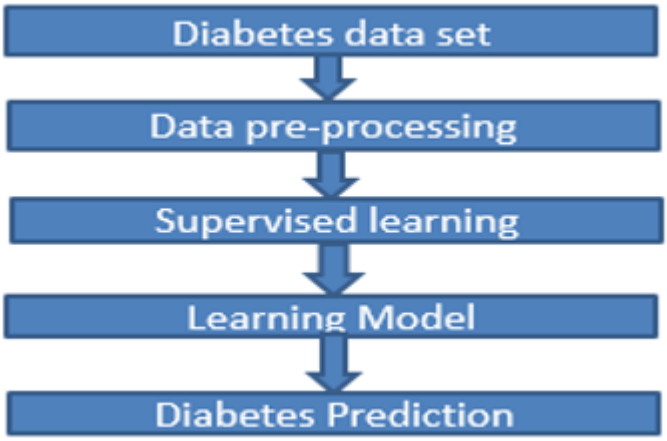


Fig. 1. Flow chart representation of diabetes prediction system

This section presents the diabetes prediction system for diabetes diagnosis. Initially the diabetes dataset is given into the data pre-processing module. The pre-processing module removes the irrelevant features from the diabetes dataset and gives the pre-processed dataset with relevant features to the machine learning algorithm. Then, the machine learning algorithm develops a learning model from the pre-processed dataset. This learning model is known as knowledge model. Furthermore, the diabetes is predicted for a person's medical report or data using the learning model.

6.1. DATASET

In data set there are nine attributes.

1	Pregnanci	Glucose	BloodPres	SkinThick	Insulin	BMI	DiabetesP	Age	Outcome
2	6	148	72	35	0	33.6	0.627	50	1
3	1	85	66	29	0	26.6	0.351	31	0
4	8	183	64	0	0	23.3	0.672	32	1
5	1	89	66	23	94	28.1	0.167	21	0
6	0	137	40	35	168	43.1	2.288	33	1
7	5	116	74	0	0	25.6	0.201	30	0
8	3	78	50	32	88	31	0.248	26	1
9	10	115	0	0	0	35.3	0.134	29	0
10	2	197	70	45	543	30.5	0.158	53	1
11	8	125	96	0	0	0	0.232	54	1
12	4	110	92	0	0	37.6	0.191	30	0
13	10	168	74	0	0	38	0.537	34	1
14	10	139	80	0	0	27.1	1.441	57	0
15	1	189	60	23	846	30.1	0.398	59	1
16	5	166	72	19	175	25.8	0.587	51	1
17	7	100	0	0	0	30	0.484	32	1
18	0	118	84	47	230	45.8	0.551	31	1
19	7	107	74	0	0	29.6	0.254	31	1
20	1	103	30	38	83	43.3	0.183	33	0

Fig. 2. Sample view of dataset

VII.EXPERIMENTAL WORK

The experiment is conducted using Jupyter notebook with the configuration of computer system 4 GB RAM, Intel core C3 6006U CPU 2.08GHz processor, Windows 10 64 – bit operating system. For the conduction of this experiment, the diabetes medical dataset has been collected from University of California, Irvine machine learning repository[9].

7.1. The data

```
Import panda as pd
Import numpy as np
Import matplotlib.pyplot as plt
```

```
From sklearn.model_selection import train_test_split
% matplotlib inline
df = pd.read_csv('diabetes.csv')
df.head (10)
```

df.info ()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 9 columns):
Pregnancies                 2000 non-null int64
Glucose                     2000 non-null int64
BloodPressure               2000 non-null int64
SkinThickness               2000 non-null int64
Insulin                     2000 non-null int64
BMI                         2000 non-null float64
DiabetesPedigreeFunction    2000 non-null float64
Age                         2000 non-null int64
Outcome                     2000 non-null int64
dtypes: float64(2), int64(7)
memory usage: 140.7 KB
```

```
df.describe ()
x = df.drop ('outcome', axis=1)
y = df ['putcome']
x_train, x_test, y_train, y_test = train_test_split (x,y, test size = 0.2, random_state = 0)
```

7.2. Prediction of diabetes using machine learning algorithms

7.2.1. Artificial neural network

In a neural network there are specifically 3 layer, input layer, hidden layer and output layer. Hidden layer is between input layer and output layer. There may be one or extra hidden layer. The layers of neural network are made up of nodes. Input layer nodes are called input nodes. Hidden layers and output layers nodes are referred to as neurodes. Every output node takes, as an enter, a weighted sum of the output from nodes inside the previous layer. Then it's far used to an activation function to the weighted input.

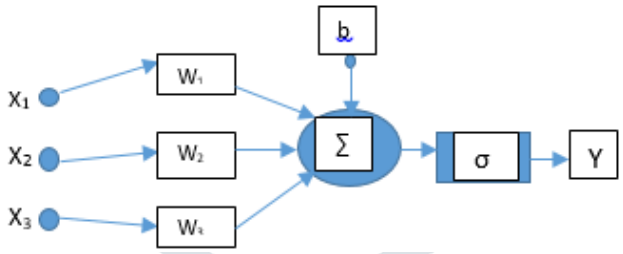


Fig. 3. . A sigmoidal unit with three inputs X = (x_1, x_2, x_3), weight vector w, bias b, sigmoid function σ and output y [10].

The output for a sigmoidal node with weight vector w and bias b on input x is :

$$\sigma\ (w.\ x+ b) = (1 +\ \exp\ (-(w\ .\ x+ b)))^{-1}$$

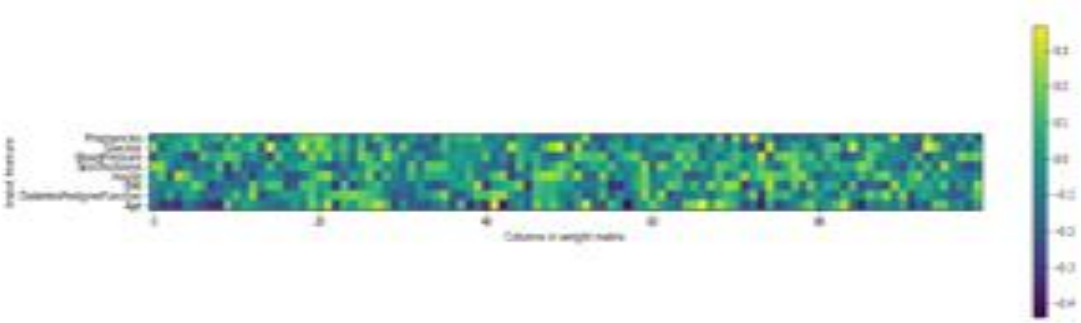
Main source code

```
from sklearn.neural_network import MLP classifier
mlp = mlpClassifier (random_state = 42)
mlp.fit (x_train, y_train)
print("training score", mlp.score(x_train, y_train))
y_pred = mlp.predict(x_test)
print ("testing score", mlp.score(x_test, y_test))
```

Training score = 0.72
Testing score = 0.70

```
plt.figure (figsize =(20,5))
plt.imshow (mlp.coefs_[0], interpolation='none',cmap = 'viridis')
```

```
plt.yticks (range(8), diabetes_features)
plt.xlabel (“columns in weight matrix”)
plt.ylabel (“input feature”)
plt.colorbar()
```



7.2.2. Support vector machine

A support vector machine is a selective thinker. It is formally known as separating hyperplane. It is a classification method that do their work by creating hyper-planes in multidimensional space that separates cases of different class labels.

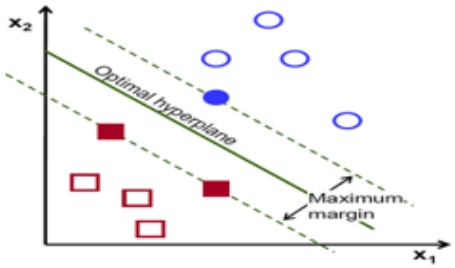


Fig. 4. Support vector machine

Methods to Compute optimal hyper-plane-

$$f(x) = h_0 + h^T x$$

This notation defines a hyper-plane
Where h is weight vector and h_0 is bias.
There are a number of different ways by which we can represent scalling of h and h_0. We can choose one

$$|h_0 + h^T x| = 1$$

Where,
x= training example, known as support vector
To find distance between a point x and a hyper-plane (h, h_0) we use :

$$\text{distance} = \frac{|h0 + h^T x|}{||h||}$$

For the canonical hyper-plane, the numerator is equal to 1 and the distance to the support vector is

$$distance_{support\ vector} = \frac{|h0 + h^T x|}{||h||} = \frac{1}{||h||}$$

Margin,

$$M = \frac{2}{||h||}$$

Main source code

```
from sklearn.svm import SVC
SVC =svc()
Svc.fit (x_train, y_train)
Print(“training score”, svc.score(x_train, y_train))
Y_pred_svc = svc.predict(x_test)
Print(“testing score”, svc.score(x_test, y_test))
Training score = 1.0
Testing score = 0.97
```

7.2.3. . K-nearest neighbors

K-nearest neighbor is used for both regression and classification problems and there is no training process for this algorithm, the entire data set is used for predicting or classifying new data.
When a new data point is given, it calculates the distance from the new data point to all others points in our data set. Then depending on the K value, it identifies the nearest neighbour in our data set,
If K = 1 then it takes the minimum distance of all points and classifies as the same class of the minimum distance data point.
If K>1 then it takes a list of K minimum distance of all data points.
For classification, it classifies the new data point based on the majority of votes in the list. For regression, it takes the average of all value in the list.

Calculate the distance from one point to another :

There are several distance metrics available, it uses one of the distance metrics

Distance functions

Euclidean $\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$

Manhattan $\sum_{i=1}^{n}|x_i - y_i|$

Minkowski $(\sum_{i=1}^{n}(|x_i - y_i|)^q)^{1/q}$

Here -
n = no. of dimensions
x = data point from data set
y = new data point (to be predicted)

Main Source code

```
from sklearn.neighbors import KNeighborsClassifier
knn = KNeighborsClassifier()
knn.fit (x_train, y_train)
print("training score",Knn.score(x_train, y_train))
y_pred_knn = knn.predict(x_test)
print(testing score", knn.score(x_test, y_test))
```

Testing score = 0.89
Training score = 0.80

7.2.4. Logistic regression

Logistic regression is used to predicts the probability of an final results which could have best two values. The prediction is based totally on using one or numerous predictors.
There is a logistic curve, that is restricted to a values among zero and 1. The curve is built using the natural logarithm of the "odds" of the target variable, in preference to the possibility.

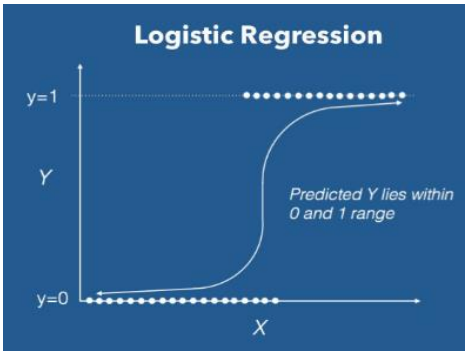


Fig. 5. Logistic regression

Logistic model

$$P = \frac{1}{1+e^{-(t_0+t_1 x)}}$$

In it, the regular t0 actions the curve left and right and the slope t1 defines the steepness of the curve. By simple transformation, the logistic regression equation may be written in terms of an odds ratio

$$\frac{p}{1-p} = \exp(t_0 + t_1 x)$$

By taking herbal go browsing both aspects, now we will write the equation in phrases of log odds that's a linear function of the predictors. The coefficient t1 is the amount the log odds modifications with a one unit change in x.

$$\ln\left(\frac{p}{1-p}\right) = t_0 + t_1 x$$

It can handle any number of numerical and/or categorical variables.

$$P = \frac{1}{1+e^{-(to+t1x1+t2x2+\cdots\ldots+tpxp)}}$$

Main Source code

```
from sklearn.inear_model import  LogisticRegression
logreg = LogisticRegression()
logreg.fit(x_train, y_train)
print("training score",logreg.score(x_train,y_train))
y_pred=logreg.predict(x_test)
print("testing score",logreg.score(x_test, y_test))
```

Training score = 0.82
Testing score = 0.76

VIII. PERFORMANCE COMPARISON AMONG DIFFERENT MACHINE LEARNING ALGORITHM

The measurement of accuracy is the ratio of truly classified samples to the total number of samples.

Accuracy=Truly classified samples / Total number of samples

Another methods for measuring performance, sensitivity and specificity are –

$$\text{Sensitivity} = \frac{true\_positive}{positive}$$

$$\text{Specificity} = \frac{true\_negative}{negative}$$

$$\text{Precision} = \frac{true\_positive}{true\ positive + false\ positive}$$

$$\text{Accuracy} = \text{sensitivity}\frac{\text{positive}}{\text{positive+negative}} + \text{specificity}\frac{negative}{positive+negative}$$

Let's see an example:

table IV

True positive = 105	False negative =10
False positive = 15	True negative = 55

Here,

n = 185

predicted yes = 120

predicted no = 65

Actual yes = 115

Actual no = 70

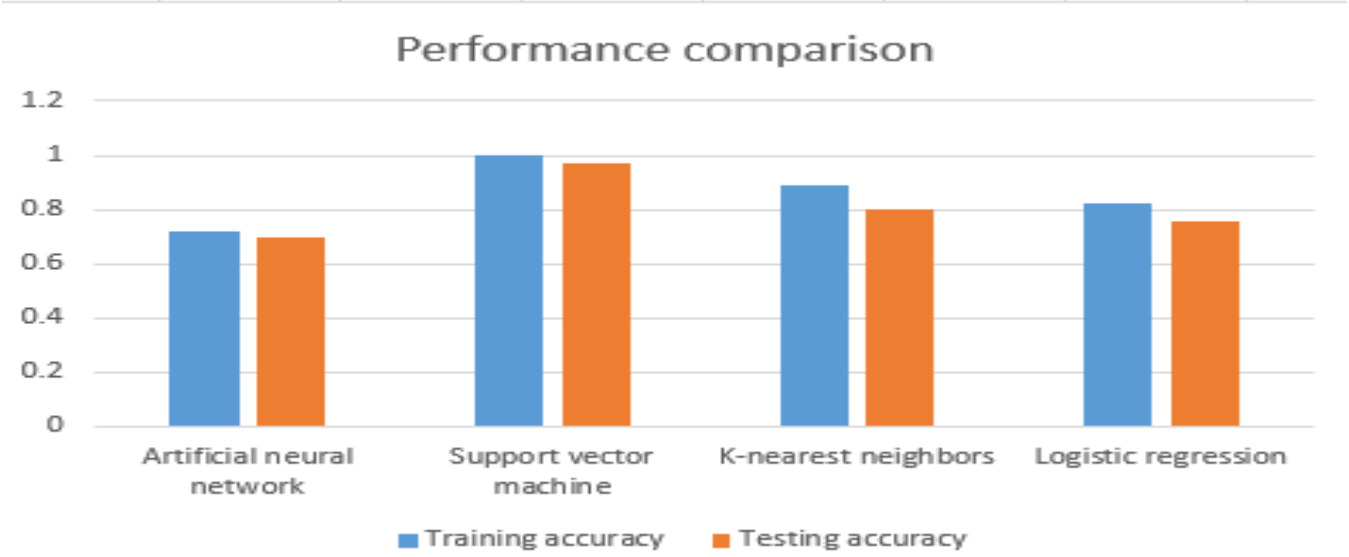
Accuracy=Truly classified samples / Total number of samples = $\frac{True\ positive+True\ negative}{Total}$

=(105 + 55)/ 185

=0.89

table V

Methods	Training Accuracy	Testing Accuracy
Artificial Neural Network	0.72	0.7
Support Vector Machine	0.1	0.97
K-nearest Neighbors	0.89	0.8
logistic Regression	0.82	0.76



IX. CONCLUSION AND FUTURE SCOPE

This paper presented a diabetes prediction system for diabetes diagnosis. In order to develop this system, the data set is collected from the University of California, Irvine repository. Different machine learning algorithm namely Artificial Neural Network, Support vector machine, K-nearest neighbors, Logistic regression are used to build the machine learning model to carry out the diagnosis of diabetes. The pre-processing technique is used to increase the accuracy of the model. From this result, it is observed that the pre-processing technique increases the accuracy of the machine learning algorithm except one case.

REFERENCES

1. Kaveeshwar, S.A., and Cornwall, J., 2014, "The current state of diabetes mellitus in India".
2. "About diabetes" WHO. Archived from the original on 31 March 2014. Retrieved 4 April 2014.
3. "Diabetes fact sheet N^0 312" WHO. October 2013. Archived from the original on 26 August 2013. Retrieved 25 March 2014.
4. Mohammed, A.K., Sateesh, K.P., Dash G.N., 2013, "A survey of data mining techniques on medical data for finding locally frequent diseases " International journal of Advanced Research in computer science and software engineering.
5. Chunhui, Z., Chengxia, Y., " Rapid model identification for online subcutaneous Glucose concentration prediction for new subjects with Type 1 Diabetes " IEEE Transactions on biomedical engineering.
6. Srinivas, K., Kavihta, R.B., Govrdhan, , A 2010 " Application of data mining techniques in healthcare and prediction of heart attacks" International journal on computer science and engineering.
7. Durairaj, M., Ranjani V., 2013 "Data mining Application in healthcare sector: A study ", International journalof scientific & technology research.
8. Nirmala Devi M., Appavu alias Balamurugan S., Swathi U.V., 2013, "An amalgam KNN to predict diabetes mellitus ", IEEE International conference of emerging trends in computing, communication and nanotechnology.
9. Lichman, M., 2013 "UCI Machine Learning Repository" [http:/archive.ics.uci.edu/ml]. Irvine, CA: University of California, school of information and computer science.
10. C. Artificial Neural Model. Retrieved July 14, 2005, from https:/commons.wikimedia.org/wiki/file: Artificial Neuron Model.png