Volume: 7 Issue: II Month of publication: February

DOI: http://doi.org/10.22214/ijraset.2019.2116

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 7 Issue II, Feb 2019- Available at www.ijraset.com

Prediction of Diabetes using Data Mining Algorithm

Chandan Kumar[1], Nanhay Singh[2], Jaspreet Singh[3]
[1]Department of Computer Science and Engineering, AIACTR
[2]Department of Computer Science and Engineering, AIACTR
[3]Department of Computer Science and Engineering, AIACTR

Abstract: Diabetes is a group of diseases where people do not produce enough insulin to meet their body's need. Insulin is a hormone secreted by pancreas. Diabetes is a leading cause of blindness, kidney failure, heart failure and stroke. The growth of the diabetic patients is increasing day by day due to the various causes such as toxic and chemical content mix with the food, bacterial and viral infections, bad diet, change in life style, environmental pollution etc. There are various methods for diagnosing diabetes based on physical and medical tests. These methods can have errors due to different uncertainties. A number of machine learning algorithm are designed to overcome these uncertainties. The main machine learning algorithms in this paper are Decision tree, Random forest, Gradient boosting and Support vector machine.

Keywords*: *Diabetes, Machine learning, prediction model, train, test score.

I. INTRODUCTION

Machine learning is a branch of artificial intelligence. It is a method of data analysis that automates analytical model building. The data analytics is a process of examining and identifying the secret patterns from large amount of data to give conclusion. In health care, this analytical process is carried out using machine learning algorithm for analyzing medical data to buld the machine learning models to carry out medical diagnosis. Diabetes is a fast growing diseases among the people even among the youngster in the whole world. Diabetes is caused by increase level of the sugar in the blood (high blood glucose). Diabetes leads various diseases such as cardiovascular complications, renal issues, retinopathy, foot ulcer etc. Diabetes is a group of diseases where people do not produce enough insulin to meet their body's need. Mainly four major types of diabetes- Type 1, Type 2, Gestational diabetes, congenital diabetes. Type 1 diabetes is an autoimmune disease. In this case, the body destroys the cells that are essential to produce insulin to absorb the sugar to produce energy. Type 2 diabetes usually affects the adults who are obese. In this type, the body resist observing insulin or fails to produce insulin. Type 2 generally occurs in the middle or aged group [6]. Gestational diabetes founds in pregnant women. During pregnancy if the blood sugar level is too high then it is considered as gestational diabetes. Congenital diabetes is very rare. It happens due to genetic defect of insulin secretion.

II. LITERATURE REVIEW

This section reviews various research works that are related to the proposed work. Chunhui Zhao et al presented a system for Subcutaneous Glucose Concentration Prediction. This model can predict Type 1 diabetes [2]. M. Durairaj and V. Ranjani discussed the machine learning techniques such as decision tree algorithm, Naïve Bayes and others to the massive volume of healthcare data. By this paper medical problems have been analysed and evaluated like heart diseases and blood pressure [3]. K. Srinivas et al developed applications of machine learning techniques in healthcare and prediction of heart attacks. In this paper medical profiles are used like patient's age, sex, blood sugar and blood pressure. Using this profile predicted the likelihood of patients getting a heart and kidney problems [4]. From this literature, it is observed that the machine learning algorithms place a significant role in knowledge discovery from the databases especially in medical diagnosis with the medical data.

III. PROPOSE WORK

A. We will try to create a standard model for diabetes prediction.
B. We will try to enhance model accuracy using boosting.
C. We will try to enhance model stability using bagging.
D. To create a model for very large dataset.

IV. TECHNICAL APPROACHES OF PREDICTIVE MODELING

There are various methods through which make prediction about future or unknown events by analyzing current or historical facts. There are various data mining and machine learning techniques through which one can make a prediction model. The most commons ones are-

A. Naïve Bayes
B. Decision Tree
C. Random Forest
D. Gradient Boosting

V. METHODOLOGY

We need data to make a prediction model. Data is the primary component that drives the analytics process. Generally there are two types of data needs to develop a prediction model

1) *Predictor Data:* It is also known as predictor variable. A predictor variable is a variable used to predict another variable. This type of data we need to make prediction. Ex – age, income, growth rate etc.
2) *Behavioral Data:* Behavioral data refers to information produced as a result of actions. It is also known as outcome data.

To make a better prediction model, the development sample need both types of data. After that an appropriate mathematical technique is applied to determine what kind of relationship between these two kinds of data. The relationship which are obtained are used in the prediction model.

VI. DIABETES PREDICTION USING MEDICAL DATA

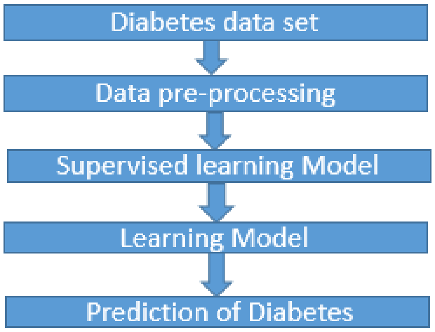


Fig.1. Flow chart representation of diabetes prediction system.

The diabetes prediction system is being presented in this section for diabetes diagnosis. There are five steps. Initially we need diabetes data-set. The diabetes data-set is given to the data pre-processing module. It removes the unrelated feature from the data-set, then it goes to the machine learning algorithm with related feature. After that, machine learning algorithm develops a prediction model from the pre-processed data-set. It is also known as knowledge model. Furthermore, the diabetes is predicted for a person's medical report or data using the knowledge model.

A. Data-Set

There are 9 attributes and 2000 number of instances in our data-set. The data-set is based on Pima Indian Diabetic set from University of California, Irvine Repository of machine learning databases [5].

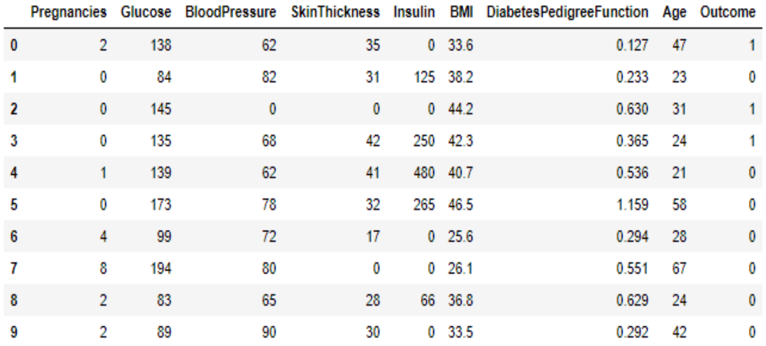
1) Number of times pregnant
2) Glucose tolerance test
3) Blood pressure
4) Triceps skin fold thickness in mm
5) 2 hour serum insulin in mu/ml
6) Body mass index
7) Diabetes pedigree function
8) Age in years
9) Outcome (1 indicate test is positive and 0 indicates test is negative)

VII.EXPERIMENTAL WORK

The experiment is conducted using Jupyter Notebook with the configuration of computer system of 4 GB RAM, Intel core C3 6006U CPU 2.0GHz 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 [5].

A. The Data

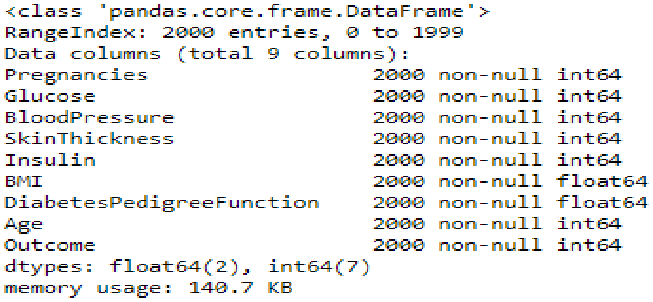
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
% matplotlib inline
import math
df = pd.read_csv(‘diabetes20.csv’)
df.head(10)
```



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	2	138	62	35	0	33.6	0.127	47	1
1	0	84	82	31	125	38.2	0.233	23	0
2	0	145	0	0	0	44.2	0.630	31	1
3	0	135	68	42	250	42.3	0.365	24	1
4	1	139	62	41	480	40.7	0.536	21	0
5	0	173	78	32	265	46.5	1.159	58	0
6	4	99	72	17	0	25.6	0.294	28	0
7	8	194	80	0	0	26.1	0.551	67	0
8	2	83	65	28	66	36.8	0.629	24	0
9	2	89	90	30	0	33.5	0.292	42	0

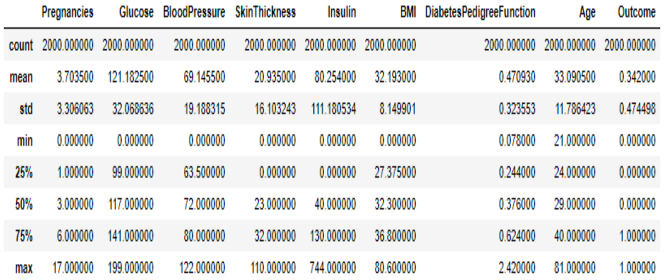
Fig.2. Sample view of data-set

```
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()
```



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000
mean	3.703500	121.182500	69.145500	20.935000	80.254000	32.193000	0.470930	33.090500	0.342000
std	3.306063	32.068636	19.188315	16.103243	111.180534	8.149901	0.323553	11.786423	0.474498
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	63.500000	0.000000	0.000000	27.375000	0.244000	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	40.000000	32.300000	0.376000	29.000000	0.000000
75%	6.000000	141.000000	80.000000	32.000000	130.000000	36.800000	0.624000	40.000000	1.000000
max	17.000000	199.000000	122.000000	110.000000	744.000000	80.600000	2.420000	81.000000	1.000000

```
x = df.drop(‘outcome’, axis = 1)
y = df[‘outcome’]
X_train, X_test, y_train, y_test = train_test_split(X,y, test_size=0.2,random_state=0)
```

B. Prediction of diabetes using machine learning algorithm

1) *Naïve Bayes:* Naïve bayes are a collection of classification algorithm based on Bayes theorem. Naïve bayes is not a single algorithm. It is a collection of algorithm where all of them share a common principle, i.e. every pair of features is independent classified of each other. Naïve bayes is a powerful algorithm for predictive modeling.

Bayes' Theorem:

$$P(A/B) = \frac{P(A).P(B/A)}{P(B)}$$

Naïve model:

$$P(B/A) = P\left(x_{i,\ldots\ldots,x_n}/A\right) = \prod_{i} P(x_i/A)$$

Attributes independent, given class

$$P\left(A/x_{1,\ldots\ldots,}x_n\right) = \acute{\alpha}P(A)\prod_{i} P(x_i/A)$$

a) Main Source Code

```
from sklearn.naive_bayes import GaussianNB
nb = GaussianNB()
nb.fit(X_train, y_train)
print("training score", nb.score(X_train, y_train))
y_pred_nb = nb.predict(X_test)
print("testing score', nb.score(X_test, y_test))
```

output:

```
Training score = 0.76
Testing score = 0.75
```

Feature importance in Naïve bayes :

It rates each feature, how important it is in making a decision.

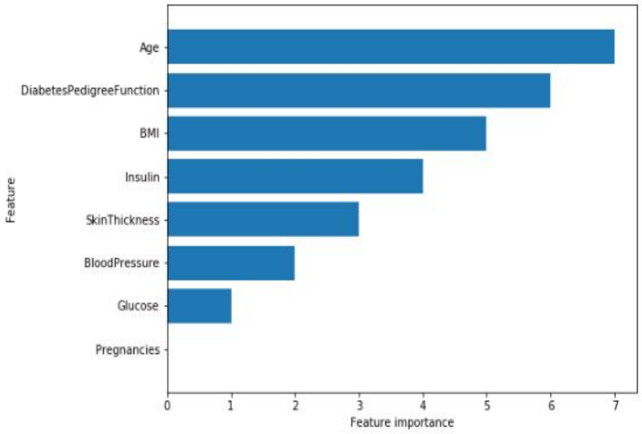


Fig.3. Feature importance

2) *Random Forest:* It is used for classification as well as regression problems. Random forest makes multiple decision trees and merges them to get a better accuracy. The large number of trees make the algorithm slow. To do accurate prediction it requires more trees, that's results in a slower model.

Let training set be [M1, M2, M3, M4] with corresponding labels [N1, N2, N3, N4], random forest create three decision trees taking input of subset

a) [M1, M2, M3]

b) [M1, M2, M4]

c) [M2, M3, M4]

At the end, it predicts based on the majority of votes from each of the decision tree made.

i) *Main Source Code*

```
from sklearn.ensemble_import RandomForestClassifier
random_forest = RandomForestClassifier(n_estimators=100)
random_forest.fit(X_train, y_train)
print("training score", random_forest.score(X_train, y_train))
y_pred = random_forest.predict(X_test)
print("testing score", random_forest.score(X_test, y_test))
output:
Testing score = 1.0
Training score = 0.995
```

Feature importance in Random Forest:

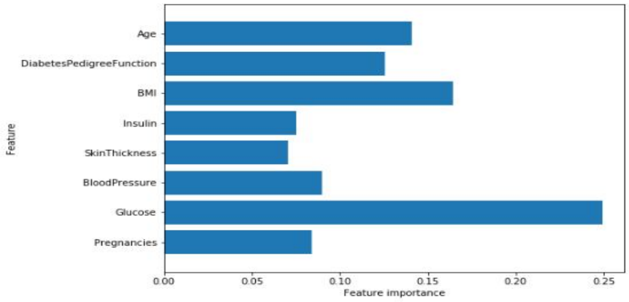


Fig.4. Feature importance

3) *Decision Tree:* Decision tree is a flow chart like tree structure. It can be binary or non-binary. In a decision tree each non-leaf node denotes a test on an attribute, every branch is an outcome of a test and each leaf contains a class label. The top node in tree is called root node [7].

The most common algorithm of decision tree is ID3. There are two mathematical tools needed to complete ID3 algorithms.

a) *Entopy:* it is used to measure important of information relative to its size.

$$H(a) = - p_+ \log_2 p_+ - p_- \log_2 p_-$$

where, a is training set

b) *Information Gain:* It is the difference between original information required and the new requirement [7].

$$Gain(a,b) = H(a) - \sum_{i \in values(b)} \frac{|\triangle b_i|}{|\triangle b|} H(b_i)$$

where, b is set of attributes

i) *Pseudo Code*

```
ID3(a, b, label, root)
initialize N as a new node
if all rows in a only have single classification c, then:
insert label c into N
return N
if a is empty, then:
insert dominant label in a  into N
return N
bestAttr is an attribute with maximum information gain in 'a'
insert attribute bestAttr into N
for vi in values of bestAttr:
insert value vi as branch of N
create viRows with rows that only contain value vi
if viRows is empty, then:
this node branch ended by a leaf with value is dominant  label in a
else:
```

```
newB = list of attributes b with bestAttr removed
        nextNode = next node
connected by this branch
nextNode = ID3(viRows, newB, label, nextNode)
return node
```

ii) Main Source Code

```
from sklearn .tree import DecisionTreeClassifier
Tree = DecisionTreeClassifier(random_state = 0)
Tree.fit (X_train, y_train)
print("training score", Tree.score(X_train, y_train))
y_pred = Tree.predict(X_test)
print("testing score", Tree.score(X_test, y_test))
```

output:

```
Training score = 1.0
Testing score = 0.98
```

Feature importance in Decision Tree

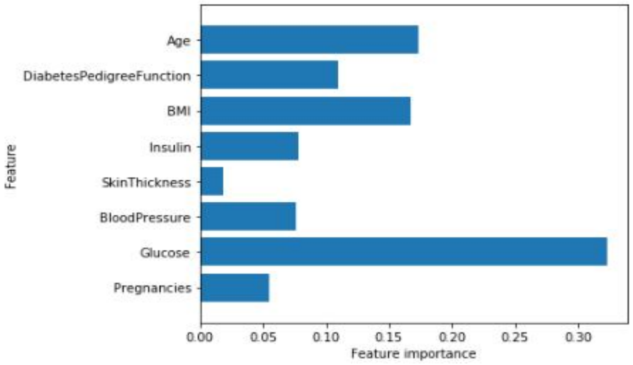


Fig.5. Feature importance

4) *Gradient Boosting:* Gradient boosting is a powerful technique for regression and classification. It ensemble weak prediction models to build an accurate model.

a) *Gradient Boosting Algorithm*

$$InitializeF_0(x) = argmin_p \sum_{i-1}^{N} L(y_i - p)$$

For m = 1 to M do:

Step 1. Compute

$$y_i = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F x_i}\right]$$

Step 2. Fit a model

$$\acute{\alpha}_m = \arg min_{\acute{\alpha}, ß} \sum_{i=1}^{N} (y - ßh(x_i; \acute{\alpha}_m))^2$$

Step 3. Choose step size

$$p_m = \arg min_p \sum_{i-1}^{N} L\left(y_i, F_m - 1(x_i) + ph(x_i; \acute{\alpha})\right)$$

Step 4. Update

$$F_m(x) = F_{m-1}(x) + p_m h(x; \acute{\alpha}_m)$$

end for

output the final regression function $F_m(x)$

b) Main Source Code

```
from sklearn.ensemble import GradientBoostingClassifier
gb = GradientBoostingClassifier(random_state = 0)
gb.fit(X_train, y_train)
print("training score", gb.score(X_train, y_train))
y_pred_gb = gb.predict(X_test)
print("testing score", gb.score(X_test, y_test))
```

Output:

```
Training score = 0.923
Testing score = 0.86
```

Feature importance in Gradient Boosting

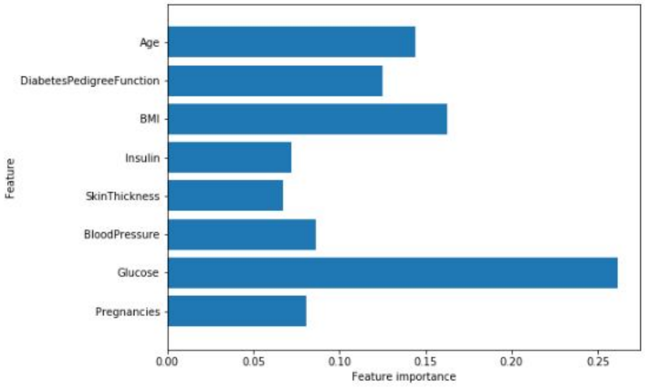


Fig.6. Feature importance

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 –

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

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

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

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

Table.1

Method	Training accuracy	Testing Accuracy
Naïve Bayes	0.76	0.75
Random Forest	1.0	0.995
Decision Tree	1.0	0.98
Gradient Boosting	0.923	0.86

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 Naïve bayes, Random forest, Decision tree, Gradient boosting 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 increase the accuracy of the machine learning algorithm except one case.

REFERENCES

JOURNAL REFERENCES

[1] C.kalaiselvi,G.m.Nasira,2014."A New Approach of Diagnosis of Diabetes and Prediction of Cancer using ANFIS", IEEE *Computing and Communicating Technologies,pp 188-190*

[2] Chunhui, Z., Chengxia, Y., 2015, "Rapid Model Identification for online Subcutaneous Glucose Concentration Prediction for new subjects with Type 1 Diabetes*", IEEE Transaction on Biomedical Engineering, 62(5), pp. 1333-1344*

[3] Durairaj, M., Ranjani, V., 2013, "Data Mining Applications in Healthcare Sector*: A study", International Journal of Scientific & Technology Research, 2(10), pp. 31-35*

[4] Srinivas, K., Kavitha, R.B., Govrdhan, A., 2010 "Application of Data Mining Techniques in Healthcare and Prediction of Heart attacks*" International Journal on Computer Science and Engineering 2(2), pp, 250-255*

[5] Lichman, M., 2013 "UCI Machine Learning Repository*" [http:/archive.ics.uci.edu/ml]. Irvine CA: University of California, School of Information and Computer Science*

[6] Kaveeshwar, S.A., *and* Cornnwall, J., 2014, "the current state of diabetes mellitus in India". *AMJ, 7(1), pp. 45-48.*

BOOK REFERENCES

[7] Jiawei Han and Micheline Kamber – Data Mining – *Concept and Techniques*; Second Edition; Elsevier Inc, 2006.