Wellness Prediction in Diabetes Mellitus Risks Via Machine Learning Classifiers

Venkatesh Saravanakumar M[1†] and Dr. M.Sabibullah [2††],

Research Scholar,
PG & Research Dept. of Computer Science,
Jamal Mohamed College (Autonomous),
Tiruchchirappalli, Tamilnadu, India

Associate Professor,
PG & Research Dept. of Computer Science,
Jamal Mohamed College (Autonomous),
Tiruchchirappalli, Tamilnadu, India.

[1†] , [2††] [Affiliated to Bharathidasan University, Tiruchchirappalli]

Abstract

The occurrence of Type 2 Diabetes Mellitus (T2DM) is hoarding globally. All kinds of Diabetes Mellitus is controlled to disrupt over 415 million grownups worldwide. It was the seventh prime cause of demise widespread with a measured 1.6 million deaths right prompted by diabetes during 2016. Over 90% of diabetes cases are T2DM, with the utmost persons having at smallest one other chronic condition in UK. In valuation of contemporary applications of Big Data (BD) to Diabetes Medicare by sighted its upcoming abilities, it is compulsory to transmit out a bottomless revision over foremost theoretical literatures. The long-term growth in medicine and, in explicit, in the field of "Diabetology", is powerfully encroached to a sequence of differences and inventions. The medical and healthcare data from varied bases like analysis and treatment tactics which assistances healthcare workers to guess the actual perceptions about the development of Diabetes Medicare measures accessible by them. Apache Spark extracts "Resilient Distributed Dataset (RDD)", a vital data structure distributed finished a cluster on machines. Machine Learning (ML) deals a note-worthy method for building elegant and automatic algorithms. ML library involving of communal ML algorithms like Support Vector Classification and Random Forest are investigated in this projected work by using Jupiter Notebook – Python code, where significant quantity of result (Accuracy) is carried out by the models.

Keywords: *Diabetes Mellitus, Medicare, Machine Learning, Big data, Spark, Jupyter Notebook, Python*

1. Introduction

Due to high growing degrees, healthcare evidences are being formed with a tall possible to convert the distribution of diabetes Medicare. BD is commencement to have an influence on diabetes care over data research. The employment of BD for repetitive Clinicalcare is still a future application. Vast dimensions of healthcare data are previously being created, and the main is connecting these to harvest an unlawful awareness. Considerable progress of effort is crucial to reach these extents.

Digital healthiness is a container, clustering composed with informatics that have the shared purposes of analysis, action, or watching of illnesses, upkeep the good health illness, and withstand for fit lifestyle. The appearance of incline of software programs with medical expedient functionalities, progressive commercial intellect data analysis tools, Artificial Intelligence (AI), cloud, cyber security are. the ground-breaking healthiness tools. These tools fit to a developing realism; they are quiet vague but these are theoretically foremost to interesting and talented situations. Workflow diagram of the projected investigation is illustrated in Fig 1.

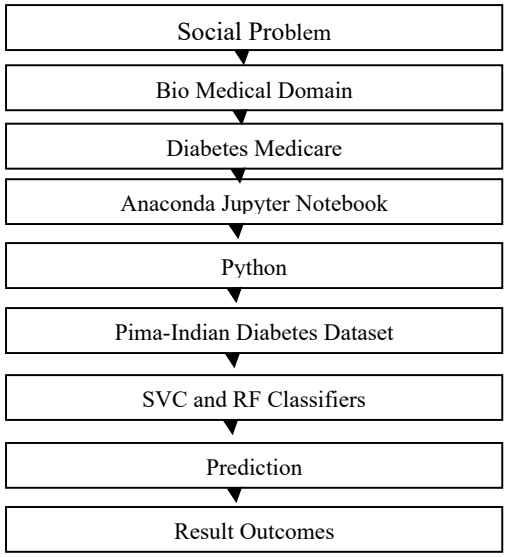


Fig 1. Workflow diagram of the projected investigation

1.1 Objectives

To develop a robust model

To deliver a good accuracy by eradicating Medical errors

To use a well structured existing medi err

1.2 Contributions

By using a well defined Diabetic datasets (Table 1), a robust and right accuracy delivered model is proposed.

Manuscript received April 5, 2022
Manuscript revised April 20, 2022
https://doi.org/10.22937/IJCSNS.2022.22.4.25

S.No.	Name of the Attribute
1	Pregnancies
2	Glucose
3	BP
4	Skin Thickness
5	Insulin
6	BMI
7	Diabetes Pedigree function
8	Age
9	Outcome

Table1. Diabetes data set attributes

2. Machine Learning Algorithms- Model Building

Machine learning is one of the indispensable and operative tools in analysing greatly multifaceted medical data. With massive quantities of medical data being engendered, there is a crucial necessity to successfully use this data to profit the medical and health care areas all across the world. Machine Learning promises to get an improved accuracy of observation and diagnosis of disease, making quick decision. Machine learning bids a note-worthy method for building an elegant and instinctive algorithm for high-dimensional bio-medical data.

2.1 Types of Learning algorithms

1. **Supervised Learning** [Teacher with Learning].Teaches the machine, how to do something, then let it use its new found knowledge to do it.
2. **Unsupervised Learning** [Teacher without Learning]. Computer can learn how to do approximately.
3. **Reinforcement Learning** (RL)
4. **Recommender Systems** (RS)

2.2 About RF and SVC

A Random Forest (RF) is a supervised learning algorithm to solve regression and classification problems. Using Support Vector Classifier (SVC) is used in the field of problems related to continuous and discrete type of data called, Classification and Regression. SVC properly sits into the training and testing data and project the best fit.

3. Literature Review

The paper [1], studied the diverse BD tools namely, Hadoop, HPCC (High Performance Computing Cluster), Storm, HBase, Grid Gain. Studied [2] the toolkits like Azure ML Studio, Amazon AWS ML, Google Cloud ML, and BigML .The survey paper [3] focused the features of Hadoop, HDFS and Map Reduce in the BD arena. Here [4], Big Data tools like Hadoop, Spark,Strom, Kafka, Flume were studied on the basis of their application area, features, and categorizes the tool and its specific areas of application. They are surveillance, IoT, environment, social media, Healthcare, Business Intelligence, Marketing and visualization. The paper [5] narrates and assess the BD processing tools, namely; Flume, HDFS, Hive, Pig, and Spark.

Analyzed [6] Apache Strom, Talend, Splice Machine, Apache Spark, Hadoop, R, Xplenty, Skytree, Lumify, Apache Cassandra, Apache SAMOA with reference to their features, various business case studies and Real World Business deployment**.** This survey paper [7], talks about a mixture of potential issues in the domain of research, and tools in BD. It is implicit that the arena of BD has their own modus operandi. Investigated [8] the impact of Hadoop and MapReduce in the financial sector. Describes an impression of BD **[9]**, environments like Spark and Hadoop, highlights the approaches to solve the issues in the Environments**.** Summarized [10] the strengths and weaknesses of Hadoop, MapReduce and Spark in connection with scalability. Hadoop and Cloud computing technologies [11] with their features were compared and listed the tools used.

4. Model Building

4.1 Algorithm 1: Diabetes Wellness Prediction using two machine learning algorithms

```
Generate training set and test set randomly.
Apply  algorithms that are used in model
SVC(),RandomForestClassifier()
Begin
{
 for(i=0; i<13; i++)
do
Model= mn[i];
Model.fit();
model.predict();    print(Accuracy(i),confusion_matrix,
classification_report);
}
End;
```

4.2 Implementation

4.2.1 Implementation / Simulation.

The present work has been experimented by using ANACONDA- JUPYTER Notebook- Python (a kind of BD Language) and following are its related components as depicted in Fig 2,3,4,5.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import os
from sklearn.ensemble import Random Forest Classifier
from sklearn.svm import SVC
from sklearn.metrics import confusion_matrix,accuracy_score,f1_score
from sklearn.metrics import classification_report
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
from scipy.stats import iqr
sns.set()
os.chdir("C:/Users/METRO/Desktop")
df=pd.read_csv("diabetes-dataset.csv")
```

Fig 2 Importing Dataset

```
x=scaled_df
y=df.Outcome
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.2,random_state=2)
model_list=[]
model_f1_score=[]
model_accuracy_score=[]
```

Fig 3 Training and Testing

```
In [46]: model_list.append('RandomForestClassifier')
         forest=RandomForestClassifier()
         forest.fit(x_train,y_train)

Out[46]: RandomForestClassifier()
```

Fig 4 Random Forest –Model-1

```
In [51]: model_list.append('SVC')
         svc=SVC()

In [52]: svc.fit(x_train,y_train)

Out[52]: SVC()
```

Fig 5 Support Vector Classification (SVC) Model- 2

4.2.2 Simulation Results and Discussion

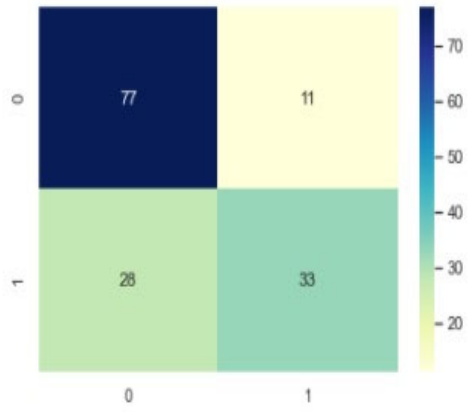
For this implementations, a Pima-Indian Dataset, where 149 samples used. Here, the 09 core and reliable attributes are used to fetch the Diabetes Medicare risk prediction.

4.3 Confusion Matrix

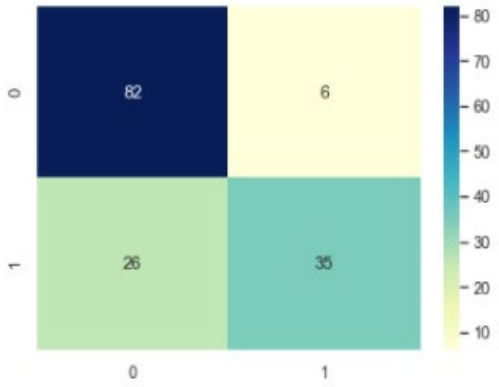
A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are defined as 1:TP: True Positive; 2.FP: False Positive; 3.FN: False Negative; 4.TN: True Negative.

TP	TN
FP	FN

4.3.1. Confusion Matrix of Model -1 - Random Forest Model (RF)



4.3.2. Confusion Matrix of Model- 2-Support Vector Classification (SVC)



4.4 Comparative Analysis

Out of two models (Both RF & SVC) , SVC has offered a very good accuracy result, is highlighted in the below Table2.

Table 2: RF and SVC Accuracy Comparison

Sl. No.	ML Classifier	Accuracy
1	RF	74 %
2	SVC	**79 %**

5. Performance Metrics(Refer Table 3)

These are listed as

1. Confusion Matrix
2. Precision
3. Recall
4. F1-score
5. Sensitivity and
6. Specificity

Table 3: Performance Metrics

Accuracy	Number of correct predictions / Total number of Predictions made
Precision= TP/(TP+FP)	It is the number of correct positive results divided by the number of positive results predicted by the classifier.
Recall=TP/(TP+FN)	It is the number of correct positive results divided by the number of *all* relevant samples. In mathematical form.
F1-Score $f1 = 2 * 1/(1/precision) + 1/recall$	It is used to measure the test's accuracy.
Sensitivity	Positive / (True Positives+False Negatives)
Specificity	True Negatives / (True Negatives+False Positives)

6. Conclusion

Healthcare outcomes in Diabetes Mellitus Medicare and treatment are coupled along with its associated cost, which is multifaceted in the incidence of comorbidities. Prognosticate the incidence of precise comorbidities may update the decision makers in this domain. So that wellness care services can be provided to the needy one. The results are compared here with its accuracies, where SVC has bestowed a better accuracy in the risk prediction of DM.

Reference

[1] J. Vijayaraj, R. Saravanan, P. Victer Paul and R. Raju, "A comprehensive survey on big data analytics tools," 2016 Online International Conference on Green Engineering and Technologies (IC-GET), 2016, pp. 1-6, doi: 10.1109/GET.2016.7916733

[2] H. Khalajzadeh, M. Abdelrazek, J. Grundy, J. Hosking and Q. He, "A Survey of Current End-User Data Analytics Tool Support," *2018 IEEE International Congress on Big Data (BigData Congress)*, 2018, pp. 41-48, doi: 10.1109/BigDataCongress.2018.00013.

[3] S. P. Menon and N. P. Hegde, "A survey of tools and applications in big data," 2015 IEEE 9th International Conference on Intelligent Systems and Control (ISCO), 2015, pp. 1-7, doi: 10.1109/ISCO.2015.7282364.

[4] B. Yadranjiaghdam, N. Pool and N. Tabrizi, "A Survey on Real-Time Big Data Analytics: Applications and Tools," 2016 International Conference on Computational Science and Computational Intelligence (CSCI), 2016, pp. 404-409, doi: 10.1109/CSCI.2016.0083

[5] S. Wadhera, D. Kamra, A. Kumar, A. Jain and V. Jain, "A systematic Review of Big data tools and application for developments," 2021 2nd International Conference on Intelligent Engineering and Management (ICIEM), 2021, pp. 561-566, doi: 10.1109/ICIEM51511.2021.9445326.

[6] S. K. Bhatt and Srinivasan, "Survey on Big Data Analytics:Domain Areas and Features," 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), 2020, pp. 254-258, doi: 10.1109/ICACCCN51052.2020.9362939.

[7]. D. P. Acharjya and Kauser Ahmed P, "A Survey on Big Data Analytics: Challenges, Open Research Issues and Tools" International Journal of Advanced Computer Science and Applications(IJACSA), 7(2), 2016. http://dx.doi.org/10.14569/IJACSA.2016.07026

[8] A. Jaiswal and P. Bagale, "A Survey on Big Data in Financial Sector," 2017 International Conference on Networking and Network Applications (NaNA), 2017, pp. 337-340, doi: 10.1109/NaNA.2017.46.

[9]. A. Jaiswal, V. K. Dwivedi and O. P. Yadav, "Big Data and its Analyzing Tools : A Perspective," 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), 2020, pp. 560-565, doi: 10.1109/ICACCS48705.2020.9074222.

[10]. M. Merrouchi, M. Skittou and T. Gadi, "Popular platforms for big data analytics: A survey," 2018 International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS), 2018, pp. 1-6, doi: 10.1109/ICECOCS.2018.8610652.

[11]Banchhor, C. O. & Srinivasu, N. (2020). Survey Of Technologies, Tools, Concepts And Issues In Big Data, international journal of scientific & technology research , VOLUME 9, ISSUE 04, pp:1901-1911,APRIL 2020,, 9.0(4.0):1901.0–1911.0.

M. Venkatesh Saravanakumar, Ph.D Research scholar in PG and Research Department of Computer Science at Jamal Mohamed College (Autonomous), Tiruchirappalli. His area of specialization includes Machine Learning, Health Big Data Analytics. He published many research papers, articles in the reputed journals

Dr. M. SABIBULLAH, currently working as Associate Professor in the PG & Research Department of Computer Science at Jamal Mohamed College (Autonomous), Tiruchirappalli, and has 23+ years of rich academic and 11+ years of research experience specializing with Machine Learning, Biomedical Big Data, Cloud Computing and Health Care Recommender applications. He published many research papers, articles; organized International Conference, National Level Technical Symposiums, Seminars, and Workshops; attended International Conferences, FDP, SDPs; delivered Radio talks in All India Radio. Acting as a reviewer in many International Conferences and chaired many National level conferences. Now, he is very keen on harnessing research interests towards hot areas of Computer Science domain like IoT, Big Data, Cloud Computing, Health Care Predictive Analytics and Data Classification algorithms. He is guiding both Ph.D and M.Phil research scholars in the line of Computer Science. He is a life member in various professional bodies.