Diabetes Mellitus prediction using different Machine Learning Algorithms and their comparison

Project report in partial fulfillment of the requirement for the award of the degree of

Bachelor of Technology

In

Computer Science & Engineering

Submitted By

Sumit Kumar Mondal	University Roll No. 12018009019161
Debalikh Chatterjee	University Roll No. 12018009019073
Atanu Mondal	University Roll No. 12018009019343
Rownak Chowdhury	University Roll No. 12018009019480
Anjali Jha	University Roll No. 12018009019154
Abhiijeet Goswami	University Roll No. 12018009019347

Under the guidance of

Prof. Stobak Dutta

&

Prof. Sumit Anand

Department of Computer Science & Engineering

UNIVERSITY OF ENGINEERING & MANAGEMENT, KOLKATA

University Area, Plot No. III – B/5, New Town, Action Area – III, Kolkata – 700160.

CERTIFICATE

This is to certify that the project titled **Diabetes Mellitus prediction using different Machine Learning Algorithms and their comparison** submitted by **Sumit Kumar Mondal (University Roll No. 12018009019161), Debalikh Chatterjee (University Roll No. 12018009019073), Atanu Mondal (University Roll No. 12018009019343), Rownak Chowdhury (University Roll No. 12018009019480)**, **Anjali Jha (University Roll No. 12018009019154) and Abhiijeet Goswami (University Roll No. 12018009019347)** students of UNIVERSITY OF ENGINEERING & MANAGEMENT, KOLKATA, in partial fulfillment of requirement for the degree of Bachelor of Computer Science & Engineering, is a bonafide work carried out by them under the supervision and guidance of Prof. Stobak Dutta & Prof. Sumit Anand during 7th Semester of academic session of 2018 - 2022. The content of this report has not been submitted to any other university or institute. I am glad to inform that the work is entirely original and its performance is found to be quite satisfactory.

_____

Prof. Stobak Dutta

Department of Computer Science & Engineering

UEM, Kolkata

_____

Prof. Sumit Anand

Department of Computer Science & Engineering

UEM, Kolkata

_____

Prof.(Dr.) Sukalyan Goswami

HOD, Department of Computer Science & Engineering

UEM, Kolkata

ACKNOWLEDGEMENT

We would like to take this opportunity to thank everyone whose cooperation and encouragement throughout the ongoing course of this project remains invaluable to us.

We are sincerely grateful to our guide Prof. Stobak Dutta and Prof. Sumit Anand of the Department of Computer Science & Engineering, UEM, Kolkata, for his wisdom, guidance and inspiration that helped us to go through with this project and take it to where it stands now.

We would also like to express our sincere gratitude to Prof.(Dr.) Sukalyan Goswami, HOD, Department of Computer Science & Engineering, UEM, Kolkata and all other departmental faculties for their ever-present assistant and encouragement.

Last but not the least, we would like to extend our warm regards to our families and peers who have kept supporting us and always had faith in our work.

Sumit Kumar Mondal

Debalikh Chatterjee

Atanu Mondal

Rownak Chowdhury

Anjali Jha

Abhiijeet Goswami

TABLE OF CONTENTS

INTRODUCTION:

Diabetes mellitus is a metabolic disease with chronic hyperglycaemia caused by many reasons which further causes increase in blood sugar. If a person has diabetes, his body either doesn't make enough insulin or can't use the insulin it makes as well as it should. When there isn't enough insulin or cells stop responding to insulin, too much blood sugar stays in the bloodstream. Over time, that can cause serious health problems, such as heart disease, vision loss, and kidney disease. It is estimated that in 2010 there were globally 285 million people (approximately 6.4% of the adult population) suffering from this disease. This number is estimated to increase to 430 million in the absence of better control or cure. An ageing population and obesity are two main reasons for the increase. Furthermore it has been shown that almost 50% of the putative diabetics are not diagnosed until 10 years after onset of the disease, hence the real prevalence of global diabetes must be astronomically high. The worldwide statistics for diabetes in the year 2013 revealed around 382 million individuals had this ailment around the world. It was the fifth leading cause of death in women and eight leading cause of death for both sexes in 2012. Higher income countries have a high probability of diabetes. In 2017, approximately 451 million adults were treated with diabetes worldwide. It is projected that in 2045, almost 693 million patients with diabetes will exist around the globe and half of the population will be undiagnosed. In addition, 850 million USD were spent on patients with diabetes in 2017.

LITERATURE SURVEY:

K. VijiyaKumar proposed random Forest algorithm for the Prediction of diabetes develop a system which can perform early prediction of diabetes for a patient with a higher accuracy by using Random Forest algorithm in machine learning technique. The proposed model gives the best results for diabetic prediction and the result showed that the prediction system is capable of predicting the diabetes disease effectively, efficiently and most importantly, instantly. Nonso Nnamoko et presented predicting diabetes onset: an ensemble supervised learning approach they used five widely used classifiers are employed for the ensembles and a meta-classifier is used to aggregate their outputs. The results are presented and compared with similar studies that used the same dataset within the literature. It is shown that by using the proposed method, diabetes onset prediction can be done with higher accuracy.

Tejas N. Joshi presented Diabetes Prediction Using Machine Learning Techniques aims to predict diabetes via three different supervised machine learning methods including: SVM, Logistic regression, KNN. This project proposes an effective technique for earlier detection of the diabetes disease. Deeraj Shetty proposed diabetes disease prediction using data mining assemble Intelligent Diabetes Disease Prediction System that gives analysis of diabetes malady utilizing diabetes patients database. In this system, they propose the use of algorithms like Bayesian and KNN (K-Nearest Neighbor) to apply on diabetes patients database and analyze them by taking various attributes of diabetes for prediction of diabetes disease. Muhammad Azeem Sarwar proposed study on prediction of diabetes using machine

learning algorithms in healthcare they applied six different machine learning algorithms Performance and accuracy of the applied algorithms is discussed and compared. Comparison of the different machine learning techniques used in this study reveals which algorithm is best suited for prediction of diabetes. Diabetes Prediction is becoming the area of interest for researchers in order to train the program to identify the patient are diabetic or not by applying proper classifier on the dataset. Based on previous research work, it has been observed that the classification process is not much improved. Hence a system is required as Diabetes Prediction is important area in computers, to handle the issues identified based on previous research.

PROBLEM STATEMENT:

Doctors rely on common knowledge for treatment. When common knowledge is lacking, studies are summarized after some number of cases have been studied. But this process takes time, whereas if machine learning is used, the patterns can be identified earlier.
For using machine learning, a huge amount of data is required. There is very limited amount of data available depending on the disease. Also, the number of samples having no diseases is very high compared to number of samples actually having the disease.

PROPOSED SOLUTION:

Goal of the paper is to investigate for model to predict dia- betes with better accuracy. We experimented with different classification and ensemble algorithms to predict diabetes. In the following, we briefly discuss the phase..

Dataset:

The dataset used in this study, is originally taken from the Pima Indian Diabetes Data set (publicly available at: UCI ML Repository).We have downloaded the file from Kaggle website. This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

Data Pre-processing: A real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models. Data pre-processing is required for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.

It involves below steps: Getting the dataset, Importing libraries , Importing datasets , Finding Missing Data , Encoding Categorical Data, Splitting dataset into training and test set , Feature scaling .

Data Cleaning: Data cleaning is one of the important parts of machine learning. It plays a significant part in building a model. It surely isn't the fanciest part of machine learning and at the same time, there aren't any hidden tricks or secrets to uncover. However, the success or failure of a project relies on proper data

cleaning. Professional data scientists usually invest a very large portion of their time in this step because of the belief that “Better data beats fancier algorithms”.

If we have a well-cleaned dataset, there are chances that we can get achieve good results with simple algorithms also, which can prove very beneficial at times especially in terms of computation when the dataset size is large.

Data Reduction: Data reduction means the reduction on certain aspects of data, typically the volume of data. The reduction can also be on other aspects such as the dimensionality of data when the data is multidimensional. Reduction on any aspect of data usually implies reduction on the volume of data

Data Transformation: Data transformation is the process of converting data from one format to another. The most common data transformations are converting raw data into a clean and usable form, converting data types, removing duplicate data, and enriching the data to benefit an organization. During the process of data transformation, an analyst will determine the structure, perform data mapping, extract the data from the original source, execute the transformation, and finally store the data in an appropriate database.

After that we use 5 types of algorithm techniques. They are-

Random Forest

From the name of this algorithm as we know that a forest is made up of trees , in the same way the random forest algorithm comprises of sub components known as the decision trees. This is a classification algorithm. It was introduced by Leo Breiman. Random Forest is a supervised learning algorithm. The final prediction about an entity is given based on the predictions by the different decision trees. That is , the final prediction is based on the prediction given by the majority of decision trees in the Random forest Algorithm just like the voting system in a democracy. It is widely believed to be the one of the best classification algorithm for high dimensional data.

Random Forest is a flexible and user-friendly machine learning algorithm that produces a great result . It is also one of the most common techniques, since it is easy to use and can be utilized for classification and regression. It is fast and easy to implement, produces highly accurate predictions and can handle a very large number of input variables without over-fitting. The forest it creates is a group of decision trees, trained most of the moment by the bagging technique and each decision tree works on different data. The general concept of the bagging technique is that the overall outcome is increased by a mixture of training designs. In plain terms: Random tree creates and merges several choices forests to make a more precise and consistent forecast. One major benefit of random forests is that they can be used for ranking as well as for regression issues.

Support Vector Machine (SVM)

Machine learning involves predicting and classifying data and to do so we employ various machine learning algorithms according to the dataset.

SVM or Support Vector Machine is a linear model for classification and regression problems. It can solve linear and non-linear problems and work well for many practical problems. The idea of SVM is simple: The algorithm creates a line or a hyperplane which separates the data into classes.

Hyperplane - A hyperplane in an n-dimensional Euclidean space is a flat, n-1 dimensional subset of that space that divides the space into two disconnected parts.

According to the SVM algorithm we find the points closest to the line from both the classes. These points are called support vectors. Now, we compute the distance between the line and the support vectors. This distance is called the margin. Our goal is to maximize the margin. The hyperplane for which the margin is maximum is the optimal hyperplane.

Logistic Regression

Logistic regression was developed by statistician David Cox in 1958. Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables. The primary difference between linear regression and logistic regression is that logistic regression's range is bounded between 0 and 1. In addition, as opposed to linear regression, logistic regression does not require a linear relationship between inputs and output variables. This is due to applying a nonlinear log transformation to the odds ratio (will be defined shortly).

Logistic function=1/(1+e−x)

It is used to predict a binary outcome (1 / 0, Yes / No, True / False) given a

set of independent variables. To represent binary / categorical outcome, we use

dummy variables.

3.4. Decision Tree

Decision Trees (DTs) are a non-parametric
supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation.
There are various algorithms in Machine learning, so choosing the best algorithm for the given dataset and problem is the main point to remember while creating a machine learning model. Below are the two reasons for using the Decision tree:
1.Decision Trees usually mimic human thinking ability while making a decision, so it is easy to understand.
2.The logic behind the decision tree can be easily understood because it shows a tree-like structure.

K-nearest neighbours (KNN)

K-nearest neighbours (KNN)

algorithm is a type of supervised ML algorithm which can be used for both classification as well as regression predictive problems. However, it is mainly used for classification predictive problems in industry. The following two properties would define KNN well −

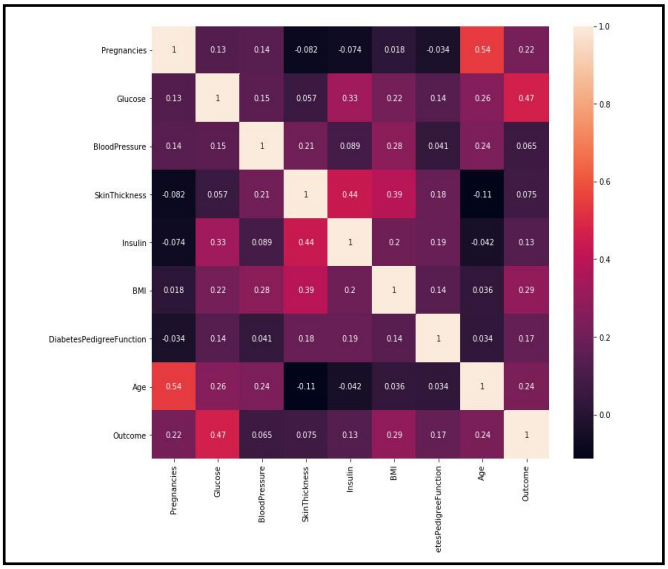
Lazy learning algorithm − KNN is a lazy learning algorithm because it does not have a specialized training phase and uses all the data for training
while classification.

Non-parametric learning algorithm − KNN is also a non-parametric learning algorithm because it doesn't assume anything about the underlying data.

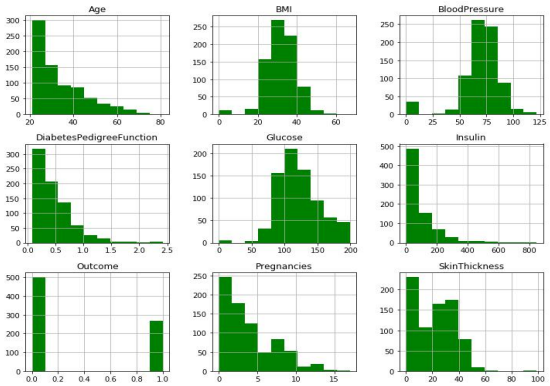
EXPERIMENTAL SETUP AND RESULT ANALYSIS:

Data Visualization:

Heatmap to show relation between different fields.



S

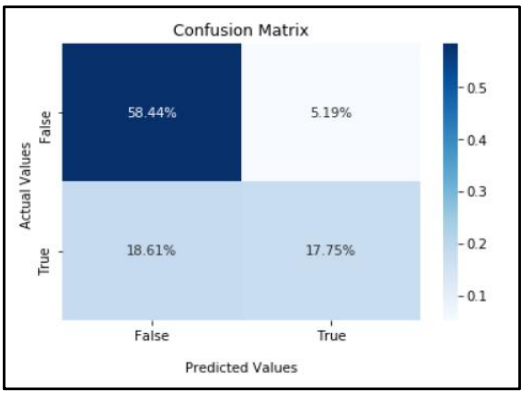


Pre-processed data visualization of different fields

Random-Forest:

	precision	recall	f1-score	support
0	0.77	0.89	0.83	147
1	0.74	0.54	0.62	84
accuracy			0.76	231
macro avg	0.75	0.71	0.72	231
weighted avg	0.76	0.76	0.75	231

performance evaluation of random fores

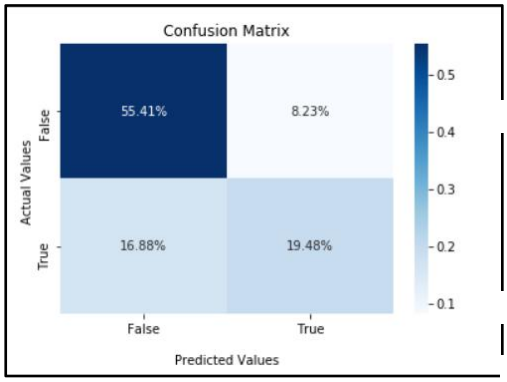


Confusion matrix of proposed model.

Support Vector Machine (SVM):

	precision	recall	f1-score	support
0	0.77	0.87	0.82	147
1	0.70	0.54	0.61	84
accuracy			0.75	231
macro avg	0.73	0.70	0.71	231
weighted avg	0.74	0.75	0.74	231

Performance evaluation of SVM



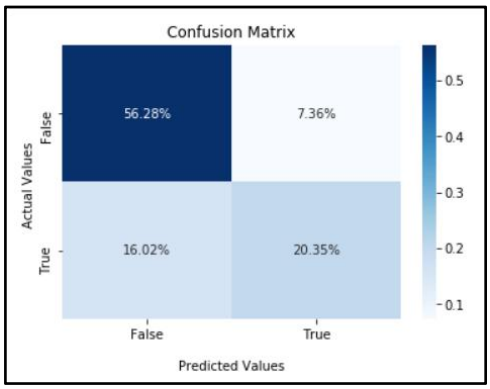
Confusion matrix of proposed model

Logistic Regression:

performance evaluation of logistic regression

	precision	recall	f1-score	support
0	0.78	0.88	0.83	147
1	0.73	0.56	0.64	84
accuracy			0.77	231
macro avg	0.76	0.72	0.73	231
weighted avg	0.76	0.77	0.76	231

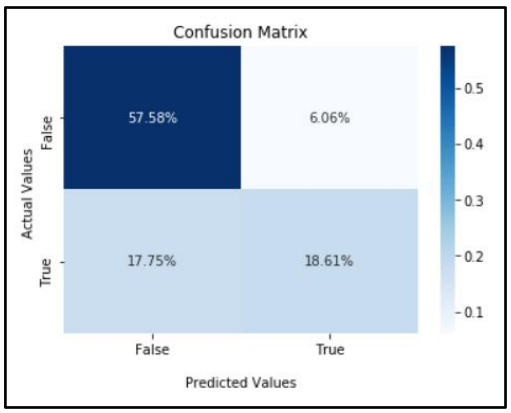
Confusion matrix of proposed model



K-nearest neighbours (KNN):

	precision	recall	f1-score	support
0	0.76	0.90	0.83	147
1	0.75	0.51	0.61	84
accuracy			0.76	231
macro avg	0.76	0.71	0.72	231
weighted avg	0.76	0.76	0.75	231

Performance evaluation of KNN

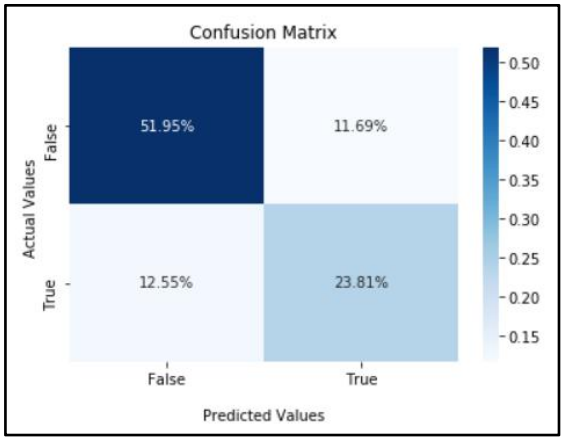


Confusion matrix of proposed model

Decision Tree:

	precision	recall	f1-score	support
0	0.81	0.82	0.81	147
1	0.67	0.65	0.66	84
accuracy			0.76	231
macro avg	0.74	0.74	0.74	231
weighted avg	0.76	0.76	0.76	231

Performance evaluation of decision tree



Confusion matrix of proposed model

CONCLUSION & FUTURE SCOPE:

Different classification algorithms were applied on our dataset, and results for all techniques were slightly different as the working criteria of each algorithm is different. The results were evaluated on the basis of accuracy. The accuracy of models was predicted with the help of a confusion matrix. First, the random forest algorithm was applied. Experiments were done to tune the model with respect to the number of decision tress and the maximum depth of the decision trees. We obtained a best accuracy of 76.19%. After the random forest algorithm, the SVM was applied to obtain better results. It divided the dataset into two classes on both side of hyperplane and obtain an accuracy of 74.8%. Similarly other algorithms like decision tree, and KNN was applied to check the performance of model and we obtain an accuracy of 75.7% and 76.19% respectively. Logistic regression was applied after all the algorithms which uses a sigmoid function and find a best fit regression between the dataset, ultimately giving us the highest accuracy of 76.6%. All the models at the end were compared and we saw Logistic regression outperformed all.

Ascending order of accuracy of models is given above and through this we can conclude Logistic regression is the best model.

	Models	Score
2	Logistic Regression	0.766234
3	Random Forest	0.761905
5	KNN	0.761905
4	Decision Tree	0.757576
1	SVM	0.748918

Reference:

1.World Health Organization, 2021 https://www.who.int/news-room/fact-sheets/detail/diabetes.

2.National Institute of Diabetes and Kidney Diseases, 2021 https://www.niddk.nih.gov/health-information/diabetes.

3.D. Sisodia, D.S. Sisodia

Prediction of diabetes using classification algorithms

Procedia Comput. Sci., 132 (2018), pp. 1578-1585

4.American Diabetes Association, Diabetes Care, https://care.diabetesjournals.org/content/30/6/1562#:~:text=Individuals%20with%20BMI%20%E2%89%A530,life-years%20lost%20to%20diabe.

5.S. Shankaracharya

Diabetes risk prediction using machine learning: prospect and challenges

J. Bioinform., Proteom. Imaging Anal. (2017)

6.Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, H. Tang, "Predicting diabetes mellitus with machine learning techniques", 2018.

7.https://www.nhp.gov.in/world-diabetes-day-2018_pg

8.Rajadhyaksha V (2018) Managing diabetes patients in India: Is the future more bitter or less sweet? Perspect Clin Res 9(1):1–3

9.https://www.quantinsti.com/blog/machine-learning-logistic-regression-python

10.Saravananathan, K., & Velmurugan, T. (2016). Analyzing Diabetic Data using

Classification Algorithms in Data Mining. Indian Journal of Science and

Technology, 9(43).

11.Krati Saxena, D., Khan, Z., & Singh, S.(2014) Diagnosis of Diabetes Mellitus

using K Nearest Neighbor Algorithm.

12.Rani, A. S., & Jyothi, S. (2016, March). Performance analysis of classification

algorithms under different datasets. In Computing for Sustainable Global

Development (INDIACom), 2016 3rd International Conference on (pp. 1584-

1589). IEEE.

13.Pradeep, K. R., & Naveen, N. C. (2016, December). Predictive analysis of

diabetes using J48 algorithm of classification techniques. In Contemporary

Computing and Informatics (IC3I), 2016 2nd International Conference on (pp.

347-352). IEEE.

14.Dataset: https://www.kaggle.com/uciml/pima-indians-diabetes-database/version/1

15.Meng XH, Huang YX, Rao DP, Zhang Q, Liu Q (2013) Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. Kaohsiung J Med Sci 29(2):93–99

16.https://diabetesatla