GESTATIONAL DIABETES PREDICTION IN PREGNANT WOMAN

Contents

List of Figures

1. Introduction

Perhaps, Diabetes Mellitus (DM) one of the oldest diseases recognized in human. About 3000 year ago, the first case of diabetes report in Egyptian manuscript. The types of DM are clearly made in 1936 which is type 1 and type 2 (Abdulfatai B. Olokoba, 2012 Jul). In 1988, type 2 was the first time define as a factor of metabolic syndrome.

Diabetes is a disease, it deals with your blood sugar (Glucose) level and how your body uses blood sugar. Blood sugar has the vital role in your health and it is a main source of energy to make the mussels and tissues of your body. It also a main source of energy for you brain to work properly. If you have the diabetes, it means body doesn’t make enough insulin and can’t use enough insulin it makes as well as it should. There are three types of diabetes e.g. type 1, type 2 and gestational diabetes. With the time, it causes the major diseases like heart diseases, vision loss and kidney diseases.

In type 1 diabetes, pancreas (is an organ of the digestive system) make very little insulin or doesn’t make insulin. Insulin (hormone) is major source of energy that enters in cells to make energy for your body. Type 1 is less common than type 2 approximately 5-10% of the people who have diabetes (Vishwendra Singh, 2021). In type 2, pancreas makes more insulin as cells need it less and pancreas can’t keep it up to maintain it. Approximately 90-95% of the people who have the diabetes of type 2. Obesity is the most import factor to cause the type 2 diabetes. Body mass index is the major factor used to define the obesity in the medical practice (Abdulfatai B. Olokoba, 2012 Jul). Gestational diabetes occurs when pancreas doesn’t make enough insulin during the pregnancy. All the pregnant woman has the insulin resistance during the pregnancy. It occurs in 5% pregnancies. The occurrence is probable to increase as the prevalent of the overweightness continues. (Ulla Kampmann, 2015 Jul 25)

In the current phase, we are suffering from the covid-19, diabetic patients high risk acquiring the infections. A study in Wuhan china and Italy indicates that diabetes is common in covid-19 patients; the reported prevalence of diabetes among the covid-19 patients varies by region and age. (Norouzi, 2021 Jun).

Problem statement

Diabetes is one the main disease causes the death. The first step to stop the progression of diabetes, Predicting and identifying the diabetes in patients. Evaluated the supervise machine learning algorithms in identifying at-risk patients using lab and survey data and also identify the key variables within the data. Here used some variables to predict the gestational diabetes in pregnant woman.

Motivation

No cure has been found yet for the diabetes, so we can detect it and control it by our lifestyle, proper checkups and diet control. Here is the case study, have used to diagnose the diabetes properly to lead the healthy and long life with the artificial intelligence based on the patient’s data. In the case, all the patients are female and at least 21 years old from pima Indian heritage dataset.

General Description of Problem

The main focus on the gestational diabetes patients. In this case study, based on the data we can find the diabetic patients. Here used some attributes like BMI, blood pressure and glucose level etc. to detect whether this patient is diabetic or not. Here used some machine learning algorithms to analyze diabetic patients. The main purpose is to predict the whether a patient has diabetes based on diagnostic measurements.

2. Related Work

In this paper, An Dinh used the different supervise machine learning models to detect the diabetes and cardiovascular diseases. He got the on average 84% accuracy. He used lab data to develop this data-driven application to diagnose diabetes and cardiovascular in patients. (An Dinh, 2019)

This study implements the different machine learning models on the authentic reports of patients given by the diabetologists. Different model has the different accuracy of the patients. Lab and report data have the different parameters to examine HPA1c, fasting blood sugar and postprandial blood sugar, these are some parameters used to train model. (Swapnil Karkhanis, 2020)

Yunzhen Ye used different machine learning and deep learning model to predict the early diabetes in the pregnant woman at 24 28 weeks of pregnancy. He used the lab data to get the features of patients. He got the 73 different features and did the features selection from the data. He didn't exceed the 95% accuracy of the models. (Yunzhen Ye, 2020)

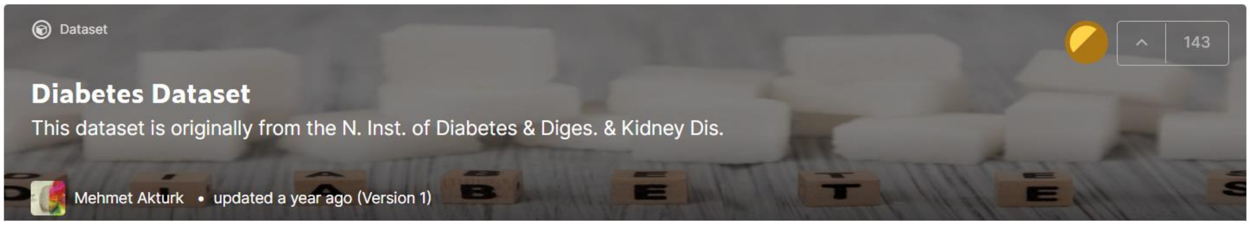
In this study, finds the gestational diabetes in patients and 15% percent woman of the world has this issue in pregnancy. He used different machine learning models to predict the gestational diabetes in pregnant woman. (Melillo, 2020)

3. Data

Data Source

In this case study, used data-set named as "Diabetes Dataset". This dataset originates from Kaggle for the non-commercial and public uses purpose only. The dataset link given below.

https://www.kaggle.com/mathchi/diabetes-data-set



Diabetes Dataset

This dataset is originally from the N. Inst. of Diabetes & Diges. & Kidney Dis.

Mehmet Akturk • updated a year ago (Version 1)

Figure 1. Data Source

This data comes from national institute of diabetes and digestive and kidney diseases. The main purpose is to predict the whether a patient has diabetes based on diagnostic measurements. These data instances from the larger dataset and many constraints were placed to select these instances. In the case, all the patients are female and at least 21 years old from pima Indian heritage.

Data Description

Here only one data file (Table) was used to analyze whether a patient is diabetic or not. Attributes ails are given below in table.

Sr. No	Diagnostic Attribute	Metadata About Attribute
1	Pregnancies	Number of times get pregnant
2	Glucose	2 hours in an oral glucose tolerance test, Plasma glucose concentration

3	Blood Pressure	Diastolic Blood pressure (mm Hg)
4	Skin Thickness	Triceps skin fold thickness (mm)
5	Insulin	Insulin, 2 hours serum (mu U/ml)
6	BMI	Body mass index (weight in kg)/ (height In m)2
7	Diabetes Pedigree Function	Diabetes Pedigree Function
8	Age	Age in years
9	Outcome	Class variable 0 or 1 (0 means not diabetic and 1 diabetic)

Exploratory Data Analysis

First of all, loaded data into working environment and showed the different aspects of data like; head & tail, dimensions of the data, columns names and data types of the columns of data etc. Checked whether the data is balanced or not. Here we can calculate the frequency of the class variables to check the imbalance data. (See fig. 2)

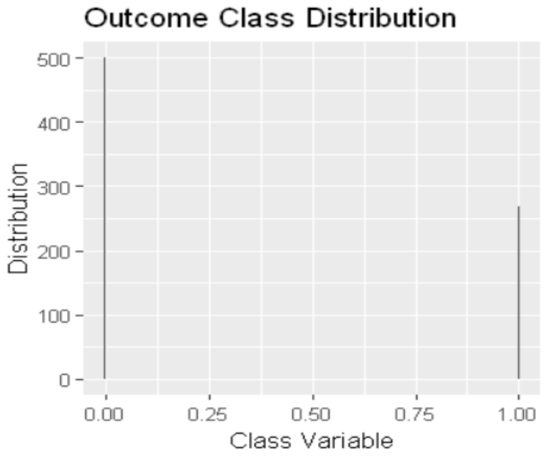
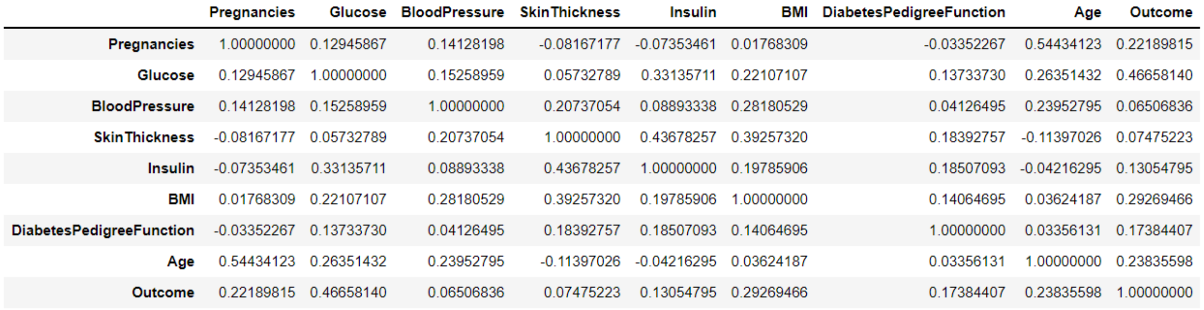


Figure 2 Frequency graph for imbalance of class variable

Statistical Analysis

First checked the standard missing values in data but can’t find any missing values, but we have some unexpected missing values in data like some columns (Glucose, BMI) have zeros. Not checked for class variable and pregnancies columns because its statistical and class variables. Impute the missing values according to the class variables with mean values of that column. Here find some statistics of the data. Correlation, description of data and summary of data are found here. Correlation is statistical association among the variables of data. It can find the linear relationship between two variables or among the variables. We can find the covariance between two variables.



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Pregnancies	1.00000000	0.12945867	0.14128198	-0.08167177	-0.07353461	0.01768309	-0.03352267	0.54434123	0.22189815
Glucose	0.12945867	1.00000000	0.15258959	0.05732789	0.33135711	0.22107107	0.13733730	0.26351432	0.46658140
BloodPressure	0.14128198	0.15258959	1.00000000	0.20737054	0.08893338	0.28180529	0.04126495	0.23952795	0.06506836
SkinThickness	-0.08167177	0.05732789	0.20737054	1.00000000	0.43678257	0.39257320	0.18392757	-0.11397026	0.07475223
Insulin	-0.07353461	0.33135711	0.08893338	0.43678257	1.00000000	0.19785906	0.18507093	-0.04216295	0.13054795
BMI	0.01768309	0.22107107	0.28180529	0.39257320	0.19785906	1.00000000	0.14064695	0.03624187	0.29269466
DiabetesPedigreeFunction	-0.03352267	0.13733730	0.04126495	0.18392757	0.18507093	0.14064695	1.00000000	0.03356131	0.17384407
Age	0.54434123	0.26351432	0.23952795	-0.11397026	-0.04216295	0.03624187	0.03356131	1.00000000	0.23835598
Outcome	0.22189815	0.46658140	0.06506836	0.07475223	0.13054795	0.29269466	0.17384407	0.23835598	1.00000000

Figure 3 Correlation among the variables

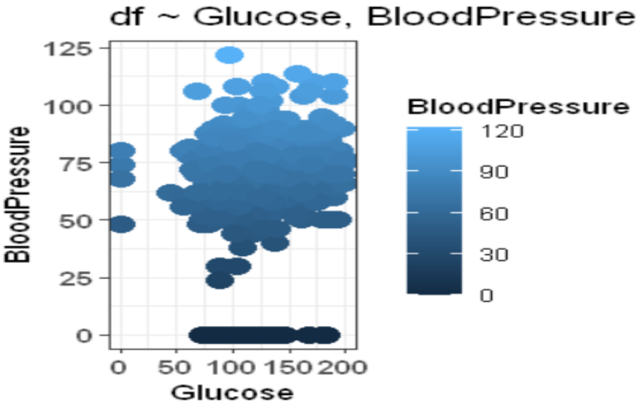


Figure 4 covariance between two variables

4. Technical Approach

Project Overview

Diabetes Mellitus (DM) one of the oldest diseases recognized in human. Diabetes is a disease, it deals with your blood sugar (Glucose) level and how your body uses blood sugar. Blood sugar has the vital role in your health and it is a main source of energy to make the mussels and tissues of your body. In type 1 diabetes, pancreas (is an organ of the digestive system) make very little insulin or doesn't make insulin. In type 2, pancreas makes more insulin as cells need it less and pancreas can't keep it up to maintain it. Gestational diabetes occurs when pancreas doesn't make enough insulin during the pregnancy.

In this project, our main focus on gestational diabetes how to prevent and control the gestational diabetes. In this case study, explored some data-driven supervise machine learning algorithms to predict the diabetes in patients. This data comes from national institute of diabetes and digestive and kidney diseases. The main purpose is to predict the whether a patient has diabetes based on diagnostic measurements. These data instances from the larger dataset and many constraints were placed to select these instances are 768. In these instances, are healthy and diabetic persons. In the case, all the patients are female and at least 21 years old from pima Indian heritage. After loading the dataset, did some exploratory data analysis and statistical analysis on data, which have used to the diabetes analytics. I have checked some basic statistics on data; find the dimensions of data, class variable (in this case Outcome) distribution, null values, imputation of null values and feature selection. Here used some machine learning models to predict and identify the diabetes in people. Machine learning model got their respective accuracies. At last machine learning models detect diabetes in patients.

My Approach

The first ever step is understanding the situation and acquiring the data through data mining techniques by converting the raw data to a suitable format for training and testing machine learning algorithms. Converted the undecipherable values to null and then fill the null values with means of column according to class variable. In this case, there is no enough preprocessing is required. Labels are signed to the class variables as 1 is diabetic and 0 is not diabetic. Its means that if the label is 1 then the patient has the diabetes and 0 has no diabetes. But in this case, we have signed the diabetic patient if has the greater than 0.5 probability and other is non-diabetic patients has less than 0.5.

In the case, all the patients are female and at least 21 years old from pima Indian heritage. Our main focus on gestational diabetes patients. After the preprocessing, split the data into training and testing dataset. In this data-driven application, down sampling is used to split the dataset into 80% / 20% training and testing dataset.

In this data-driven application, we used multiple machine learning models for the classification of diabetic patients. Training dataset that have both the class variable and class label for the category of the class variable and trained by machine learning algorithms. The algorithm can predict the label which associate with the new test set done by the model on the new class variable.

In this study, we only examine the random forest algorithm. Random forest that develops with the several random decision trees through the bagging process. Every random decision tree shows the possible results. For the final classification, calculate the average results among all the trees which are considered in the modeling. Results splits are made based on information gain.

First of all, random forest trained on default parameters. Random forest default parameters are class variable, training data and algorithm family. We got the 87% accuracy with the default parameter and mean squire root error is 0.26. Below showed the model graph to visualize the model performance. (See fig 5)

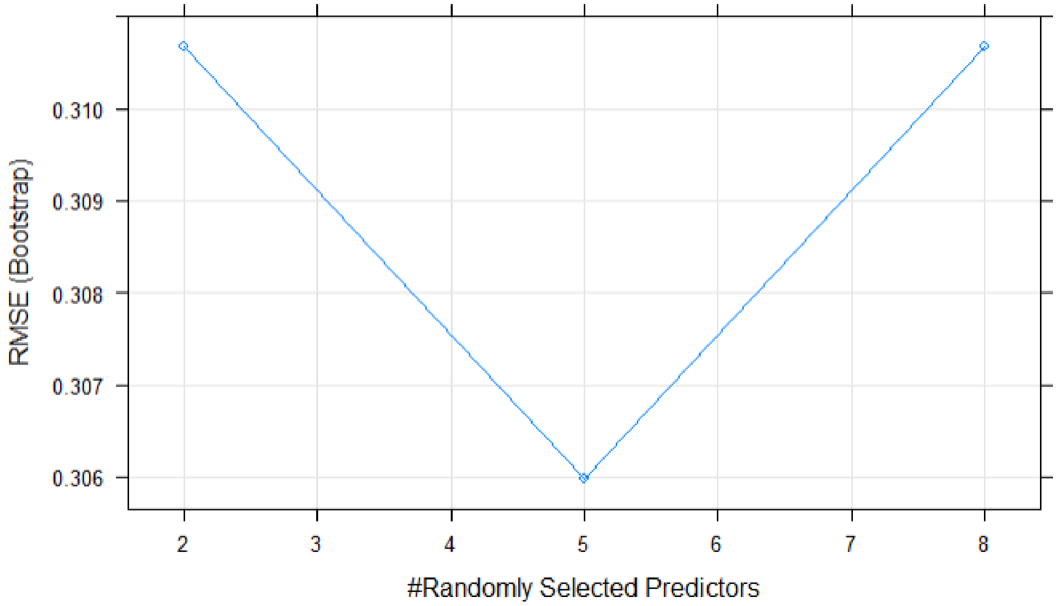


Figure 5 Model Performance with default parameters

Feature Scaling

Try to improve the accuracy and performance of the model to fine tune our model, add scaling in our model to standardize the selected features to train model on that. After fine tune the model to examine the model on new set of parameters. Somehow, we cannot successful to get enough accuracy and performance of model. See fig. 6

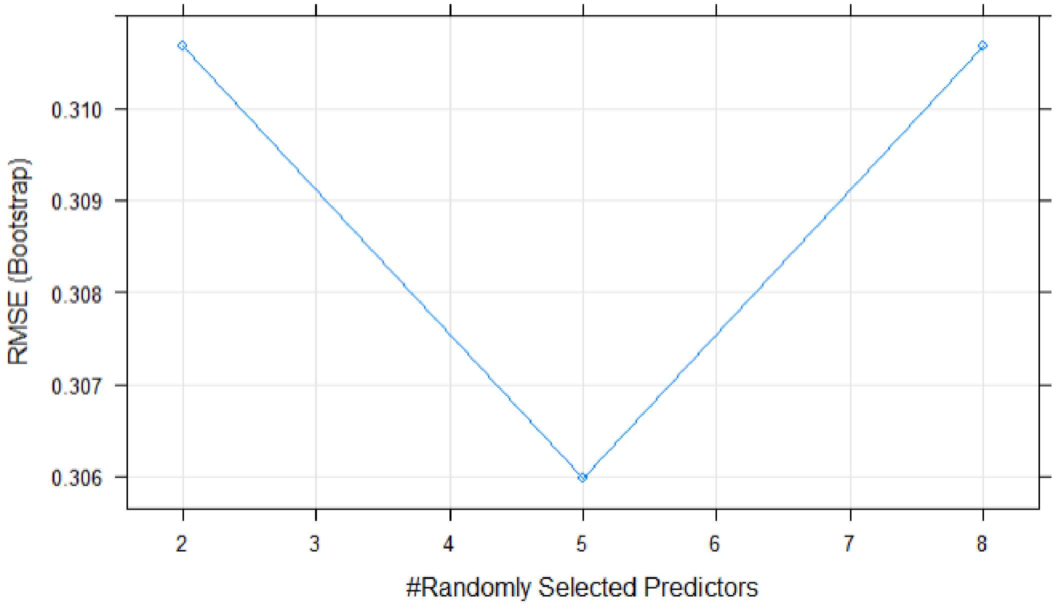


Figure 6 Model Performance with scaling

K-fold Cross Validation

Continue to improve the performance of the model, here I have added k-fold cross validation to our model. In this divide the dataset into k folds and hold one-fold from them and trained the model remaining k-1 folds and calculated the MSE on the test data. In our case, used 10-fold cross validation with CV method. See fig. 7

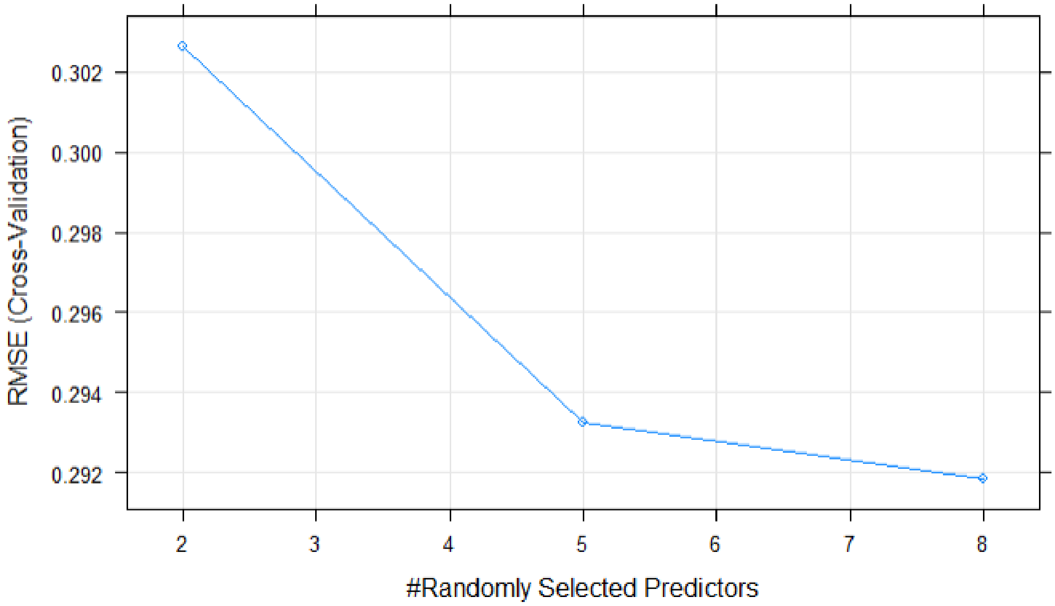
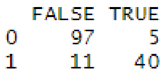


Figure 7 Model Performance with K-fold cross validation

With the cross validation, got the 89% accuracy of the model on new test data. Hence, we have increased the performance of the model using cross validation method.

5. Test and Evaluation

- After training the model, evaluate the model by confusion matrix, accuracy of the model and ROC curve of the model to ensure that model. I have tested the model on new or unseen dataset to check the validity of the model. On the training it showed the 99% accuracy and on the testing data it showed the 89% accuracy.
- To evaluate the model, found the accuracy, confusion matrix, ROC curve, root mean square error and correlation between actual and predicted, these coefficients showed the good model performance.
- Baseline of this case study is the different approaches used to predict the diabetes with machine learning and with lab data. It will compare the accuracies of different machine learning models but in my case, I used only one machine learning model to predict the gestational diabetes in pregnant women.
- Here are some metrics and results that showed the performance of the model on test data.
    1. Confusion matrix



```
   FALSE TRUE
0     97    5
1     11   40
```

    2. Accuracy

89.54 %

    3. ROC Curve

It is performance measurement of the machine learning model. Roc curve gives us trade-off between the true positive rate and false positive rate.

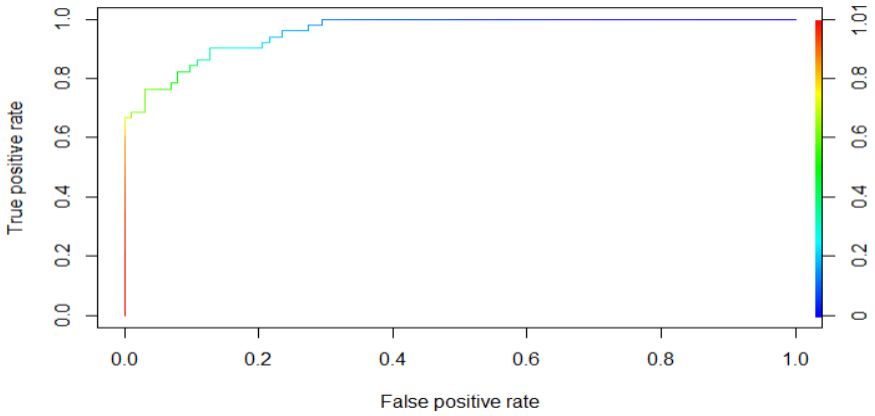


Figure 8 ROC Curve

4. Correlation between actual and predicted values
   68.89%
5. Root mean square error (how good fit the line to the data point)
   0.27 %

- The focus of this study is to find the diabetic and non-diabetic patients but the main purpose of the project is finding the gestational diabetes in the pregnant woman how much chance they have in their pregnancy. The above performance of the model shows that it will enough good to find the chance of diabetes in pregnant woman. Somehow, we are able to estimate the diabetes in pregnant woman.
- First get the desire dataset, do preprocessing on data if needs, scale the data, split the data into training and testing dataset, fit the model on data and then evaluate the model by doing training and testing analysis on it.

Future Work

It will be improving by doing the dataset well prepared for the modeling. In the data description section, described that our dataset is imbalanced and almost 500 and 268 class variables partitions. It means that 500 non-diabetic records and 268 diabetic records in the dataset. Someone will improve the model performance by adding more accurate lab data into training and testing dataset. It can be done a web portal type application or mobile application to generate the questionnaire to predict the gestational diabetes.

Conclusion

In our study, that develop a machine learning model to predict the patients who have the diabetes or not by the model performance. Here we are not going to compare the different machine learning model by their performance. We used a single machine learning model to predict the diabetes patients by fine tune the model with scaling and K-fold cross validation. This model showed the accuracy 89.54% on new test data and 99% on training set.

References

Abdulfatai B. Olokoba, O. A. (2012 Jul). Type 2 Diabetes Mellitus: A Review of Current Trends. *Oman medical journal vol. 27,4 (2012): 269-73. doi:10.5001/omj.2012.68*, 8.

An Dinh, S. M. (2019). A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. *BMC medical informatics and decision making, 19(1)* (p. 13). USA: BMC .

Melillo, G. (2020). Machine Learning Predicts Incidence of Gestational Diabetes. *AJMC*, 10.

Norouzi, M. (2021 Jun). Type-2 Diabetes as a Risk Factor for Severe COVID-19 Infection. *Microorganisms;9(6):1211. doi: 10.3390/microorganisms9061211. PMID: 34205044; PMCID: PMC8229474.*, 12.

Swapnil Karkhanis, M. W. (2020). Detection and Risk Analysis of Diabetes Using Machine Learning. *International Research Journal of Engineering and Technology (IRJET)* (p. 5). Maharashtra, India: IRJET.

Ulla Kampmann, L. R. (2015 Jul 25). Gestational diabetes: A clinical update. *World J Diabetes;6(8):1065-72. doi: 10.4239/wjd.v6.i8.1065. PMID: 26240703; PMCID: PMC4515446.*, 7.

Vishwendra Singh, K. G. (2021). Effect of an Oral Health Preventive Protocol on Salivary Parameters and Gingival Health of Children with Type 1 Diabetes. *Int J Clin Pediatr Dent*, 6.

Yunzhen Ye, Y. X. (2020). Comparison of Machine Learning Methods and Conventional Logistic Regressions for Predicting Gestational Diabetes Using Routine Clinical Data: A Retrospective Cohort Study. *Hindawi* (p. 10). Shanghai, china: Journal of Diabetes Research.