National College of Ireland

Project Submission Sheet – 2019/2020

School of Computing

Student Name:	Victory Chimamaka Uwaoma		
Student ID:	x19210931		
Programme:	Data Analytics	**Year:**	2020
Module:	Domain Application of Predictive Analysis		
Lecturer:	Vikas Sahni		
Submission Due Date:	23rd August, 2020		
Project Title:	A Predictive Analysis on Pima Indians Diabetes using machine learning		
Word Count:	4189		

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature: Victory Chimamaka Uwaoma

Date: 23/08/2020

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

A Predictive Analysis on Pima Indians Diabetes using Machine Learning

Victory Chimamaka Uwaoma
Msc. Data Analytics
National College of Ireland
Dublin, Ireland
x19210931

***Abstract*—The healthcare domain is an area that requires urgent attention, efficiency and reliability as it deals with the lives of people. Diabetes is an irreversibly deadly disease that affects millions of people. This is as a result of excessive glucose which the body cannot process leading to damaged nerves, eye disease, kidney problems or eventually death. There are different types and stages of diabetes which includes Type I diabetes, Type II diabetes and Prediabetes and Gestational diabetes, all of them having their own consequences. The approach will dramatically reduce the risk of disease by finding out from a scientific archive a simple and understandable pattern for diabetes. Machine learning has been used to gain attention in the healthcare domain. Research has been done in this area in order to get accurate prediction using machine learning algorithms. Random forest a machine learning algorithm uses lots of decision trees to train the datasets, applying the weight of each tree to get the final outcome. This experiment uses the PIMA Indian dataset which contains over 768 data samples and 8 variables for classifying whether or not a person has diabetes or not.**

This model can be implemented in the health care system as an automated predictive system for diabetes as this will help Endocrinologist in the healthcare domain in predicting diabetes fast and easy as it requires little or no human intervention. Therefore, this paper used Random forest algorithm based on previous implementation by other researchers to predict whether or not a person has diabetes as this has been tested in other diabetes datasets and has seen to provide efficient results. The end result shows that Random Forest achieved an accuracy of 77.49%, sensitivity and specificity of 0.8355 0.6582 respectively which outperform the state-of-the-art result.

Keywords—healthcare, diabetes, machine learning, random forest algorithm, prediction

I. Introduction

Diabetes is a deadly disease amongst several other diseases that does not have a cure yet. It costs a great deal of money annually to carter for people living with diabetes [1]. According to the reports from the World Health Organization (WHO), about 4.8% people globally suffer from diabetic retinopathy, an eye disease that is caused by diabetes [2]. It is important to note that with the use of machine learning algorithms for predictive analysis such as diabetes will help enhance the prognosis of the people. The use of predictive classification in medical diagnosis has got a tremendous improvement over the years due to severe research activities in this area.

This research work aims to use machine learning to model diabetes detection within PIMA Indians. Random forest is used to classify and identify the factors that can be used in diabetes prediction.

1.1 Research Goal

This research project is aimed at predicting diabetes in Pima Indians with the help of machine learning techniques. This will predict if a person is likely to have diabetes based on some factors such as how many times is the person pregnant, the person body mass index, the age of the patient. All these factors will be used in training the model and will therefore help Endocrinologist in the Health domain to determine when an eye disease is present in a patient.

1.2 Structure of Thesis

The remaining part of this paper is organized according. **Section Two – Previous Applicable Techniques:** This section talks able the Previous applicable techniques that have been done in the area of predicting diabetes.

Section Three – Methodology: This section is on the Methodology which describes the approach and implemented algorithm which was carried out during the building of the model. It will contain the proposed methodological approach this project will take in achieving its goal and also the design of the system as well as the requirements and this will be done using relevant diagrams, tables and figures.

Section Four – **Results and Discussion**: This is the Results of the implemented algorithms and the feedback on the model.

Section Five – **Other relevant features:** Other relevant features discovered while carrying out the project will be seen in Section V.

Section Six – References: This section will contain the cited references.

II. Previous applicable techniques

Different researches have been carried out in the healthcare area with the aim to provide better and improved technology as well as efficient results and also in understanding the data generated in that field. Machine learning algorithms has been seen as one of the such to be used to produce efficient results especially in the area of predicting diabetes. Several algorithms have been used in recognizing problems and finding gaps in current methodologies to develop a predictive model that can diagnose and treat diabetes. Healthcare domain researchers have used various machine learning techniques to predict diabetes among patients. The existing literature on the baseline approach

which will be used to evaluate the data in our analysis is discussed in this chapter.

According to [3] who used five different machine learning model namely k-nearest neighbour (k-NN), multifactor dimensionality reduction (MDR), radial basis function (RBF) kernel support vector machine, linear kernel support vector machine (SVM-linear), artificial neural network (ANN) and to early detect whether or not a patient is diabetic or not. This was modelled on the PIMA Indian diabetes dataset which has a total instance of 768 samples which can be categorized into two classes; diabetes or no diabetes. The experiment which partition the data into 70% training and 30% testing, feature extraction and hyperparameter tuning was done as well as applying tenfold cross validation to prevent overfitting and underfitting of the model. produced significant results on the different models. The models were all evaluated using evaluation metrics such area under curve (AUC), F1 score, recall and precision. At the end of the experiment, the SVM-linear produced an accuracy of 0.89 and precision of 0.88 unlike the other models.

The authors in [4] used J48 decision tree to predict diabetes in human from their hospital records in order to prevent other complicated diseases like blindness, damaged blood vessels, kidney diseases. Using data mining tool such as WEKA on the PIMA Indian dataset, [4] was able to build a model that could automatically diagnose if a patient is with or without diabetes. The end result showed that the model produced efficient results which can be used to model diabetes.

[5] also using the WEKA tool to predict diabetes with Random Forest, SVM, Naive Bayes, and Simple CART algorithms. They were able to build four predictive diabetes models in the healthcare domain. The experiment involved training and testing the models and using performance metric was done to check the model. This metrics include accuracy, recall, F-Measure and precision. The study revealed that SVM outperformed all other models with an increased accuracy.

In predicting different types of diabetes in women, [6] proposed comparative machine learning algorithms namely Logistic Regression, SVM, Naïve Bayes, Decision Trees and KNN for diagnosing diabetes mellitus. The model was trained and tested. Using KFold and cross validation, the final result from the experiment gave logistics regression an accuracy of 81.1% after testing the model.

Artificial neural network (ANN) has been used to predict diabetes. Researchers in [1] centered their work on the factors that cause diabetes. This model used the JNN tool environment to model the data which contained 9 attributes and 1004 tests which was trained, validated and tested. During testing, the ANN model achieved an accuracy of 87.3% when trained on 158,000 epochs, with a training sample of 767, which was then validated on 237 samples.

A machine learning approach was used in finding patterns and diagnosing diabetes as proposed by [7]. The authors made use of AdaBoost algorithm with Decision Stump as base classifier, Naive bayes, Decision tree and SVM. The dataset used was gotten from the University of California, Irvine (UCI) source which consist of 768 test samples and 9 factors. The data was pre-processed before been trained. The AdaBoost algorithm with Decision Stump as base classifier produced significant results with accuracy of 80.7% better than the others with reduced error rate.

Four machine learning algorithms namely C4.5 decision tree, naive bayes, support vector machine and k-nearest neighbor were implemented in [8] in order to early predict diabetes mellitus by discovering the various risks factors that causes diabetes mellitus. This approach made use of dataset gotten from the diagnostic of Medical Centre Chittagong (MCC), Bangladesh. The dataset contained several risk factors taken from over 200 patients with diabetes mellitus that were modelled. The data was split into training and testing. The test results reveal that [9] the decision tree C4.5 has attained better performance in comparison to other approaches of machine learning.

In order to tackle diabetes mellitus, [10] proposed machine learnings such as decision tree (DT), Random forest (RF) and neural network. Using principal component analysis (PCA) and minimum redundancy maximum relevance (mRMR), the authors were able to perform dimensionality reduction in order to reduce the cost and computation time. Using two datasets, one by randomly selecting samples of healthy and diabetic patients, the samples amounted to 68994 gotten from physical test from hospital in Luzhou and the other from PIMA Indians diabetes dataset which contains samples of females from the age of 21 and has 8 variables and 768 samples. The data was trained using WEKA tool for the RF and DT classifier and MATLAB for neural network and tested. The result showed that random forest performed the best amongst other classifiers with 80.8% on Luzhou dataset and 77.2% on PIMA dataset.

The research in [11] focuses on predicting the risk of diabetes and building a predictive model for diagnosing whether or not a person has diabetes or not. Using a public dataset gotten from National Inpatient Sample (NIS) available by Healthcare Cost and Utilization Project (HCUP), the data was trained using random forest, SVM, boosting, and bagging to predict 8 incurable diseases. The experimental result showed that random forest was a better classifier with an area under the curve (AUC) as 89.05%.

In [12] the authors proposed the use of random forest and other machine learning algorithms in predicting type II diabetes. The data was pre-processed to improve the accuracy when building an efficient model. The dataset was obtained from University of Virginia, the Medical department contains 403 samples with 19 variables. The experiment was divided into 4 which trained and tested the data and the performance metrics showed that random forest was the best with 84.19% accuracy.

The use of four predictive model to determine if a person has diabetes was proposed by [13]. The experiment made use of the dataset from UCI with 8 variables and 1 predicting variable. Using the WEKA tool to model the data. The first experiment was not done with pre-processing and the result showed J48 classifier had a highest accuracy of 73.82% but in the second experiment for modelling a predictive model for diabetes, random forest and K-NN (k=1) achieved the same results of 100% accuracy, specificity and sensitivity.

[14] used five classification algorithms to classify the risk factors of diabetes mellitus using a web application. The algorithms which includes ANN, logistics regression, naive bayes, random forest and decision tree were trained using data gotten from Primary Care Units (PCU) in Sawanpracharak. The dataset had a little over 30,000 samples with 12 variables. The experiments modelled were 13 with the inclusion of

bagging and boosting. The random forest proved to be more efficient amongst all the other classifiers with an accuracy of 85.558%.

Using optimal feature selection to early classify diabetes mellitus in [15] using predictive analysis, the researchers proposed a method with the use of five classification algorithms. The aim of the analysis was to determine the features that cause diabetes. The dataset used contains 15 attributes and 2500 samples was gotten from UCI data repository. During the course of training the model, 11 features were selected and trained using the predictive classifiers. The model was also tested. Performance metrics such as accuracy was used in evaluating the model. The decision tree was seen to have an accuracy of 98.20%

After carrying out extensive literature in the area of diabetes, in order to achieve the aim of this project, the Random forest will be implemented in other to predict whether or not a person is diabetic as this has been seen as one of the classification algorithms that has produced significant result according to [10].

III. Methodology

The machine learning techniques used based on all other techniques that has been used in this area is the Random forest. Random forest algorithm is a classifier that deals well with both classification and regression problem. According to [10] who used Random forest algorithm in predicting diabetics and achieved significant result.

This project will adopt the approach of Knowledge Discovery in Databases (KDD), which will be applied to the Pima diabetes database in predicting if a person is diabetic or not. This methodology was chosen based on previous applicable techniques that has used it and has proven to produce reliable results.

The following steps outlined is according to the KDD methodology:

a. Data sourcing
b. Data pre-processing
c. Data selection
d. Implementation of Algorithm
e. Result

a. Data sourcing

The Pima diabete dataset is gotten from GitHub https://raw.githubusercontent.com/jbrownlee/Datasets/master/pima-indians-diabetes.csv. The dataset is owned by the National Institute of Diabetes and Digestive and Kidney Diseases and consists of information on diabetic samples collected from the Indian women of Pima. It comprises of 8 attributes and 768 samples. Below shows more information about the dataset.

Table 1: Pima Indians Diabetes dataset

S/N	Variable	Type	Description
1	Pregnancies	Numeric	Number of times pregnant
2	Glucose	Numeric	Plasma glucose concentration a 2-hours in an oral glucose tolerance test
3	Blood Pressure	Numeric	Diastolic blood pressure (mm Hg)
4	Skin Thickness	Numeric	Triceps skin fold thickness (mm)
5	Insulin	Numeric	2-Hour serum insulin (mu U/ml)
6	BMI	Numeric	Body Mass Index (weight in kg) / (height in m^2)
7	Diabetes Pedigree Function	Numeric	Diabetes Pedigree function (DPF)
8	Age	Numeric	Ages (years)
9	Outcome	Categorical	Class variable (0 or 1) 1– diabetic 0– non diabetic

b. Data pre-processing

In this phase, the sourced data was imported in RStudio and was cleaned and transformed according for analysis. Firstly, the necessary libraries were imported. The dataset was missing a variable header which was programmatically encoded, then the data was discovered to have NA (not available) valves which could not be removed due to the small nature of the dataset but was programmatically replaced by the mean of each variables.

c. Data selection

At this point, only the variables important for the study have been chosen. Selection was done using a heatmap. The heatmap is a graphical representation that is used to visualize the data and show relationships between the independent variables and its dependent variable.

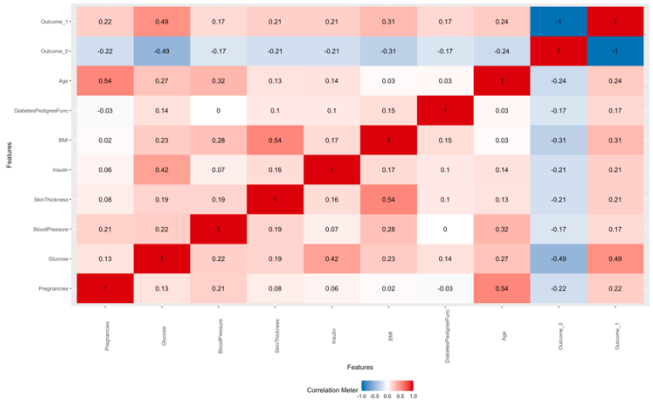

Figure 1: Heatmap of the diabetes dataset

From the heatmap above in figure 1, it shows that there was correlation between BMI and Skin thickness, Age and Pregnancies, therefore, they were not removed due to the small nature of the dataset as this will affect the analysis from producing optimal result for the model.

d. Implementation of Algorithm

The algorithm used in the prediction of the Pima dataset wass selected based on the review done by [10] who used similar dataset and achieved significant results. Random forest algorithm was used on the Pima Indian diabetes dataset. Upon loading the necessary libraries, the data dimension and structure were checked to ensure that the data was entered correctly. Pre-processing and visualization was done on the data in other to have a better view of the relevant information to be used when developing the predictive model for the problem. The random forest classifier was used to fit the diabetes data. The data which was divided into 70 per cent for training and 30 per cent for testing the research study.

e. Result

The aim of this analysis was to critically analyse the Pima Indian diabetes dataset looking for patterns that can tell if a person is diabetic or not and analyse based on the factors present in the datasets. Using random forest classifier, the model was able to perform significantly more than the state-of-the-art model. It had an accuracy of 77%. The random forest classifier used a mtry of 6 which equals number of random samples the predictors will split when creating the tree models. The random forest plot in figure 2 below shows that the data set is split into two child nodes at the parent node Glucose. If the glucose present is greater than 139, it means the patients has glucose in their system and goes further to check for other factors and if less than or equal to 139, the patient potentially has glucose and goes further to also split the node based on other factors.

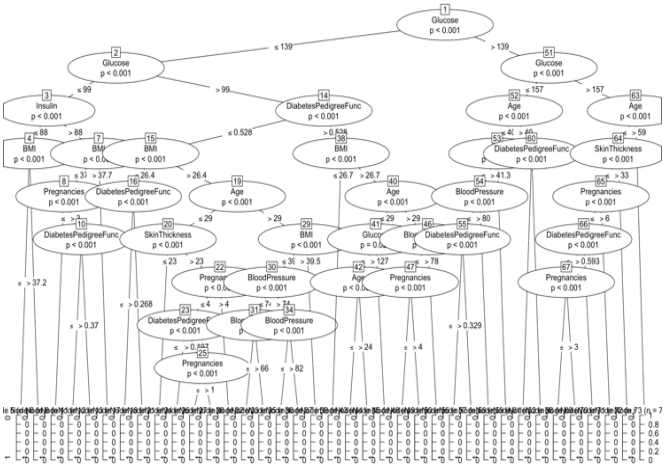

Figure 2: Random forest Tree

IV. Results and Discussion

In this study we used the PIMA Indian diabetes created by the National Institute of Diabetes and Digestive and Kidney Diseases. At the end of the analyses, promising results were gotten that performed better than the state-of-the-art model [10]. The model got an accuracy of 77.49% and sensitivity 83% also known as recall shows us that the model can predict positives correctly and 65% specificity can predict negative as negative. The table 2 below shows the model summary of the random forest.

Table 2: Summary of the model

Accuracy	Sensitivity	Specificity
0.7749	0.8355	0.6582

In the business sense, the medical diagnosis system achieved an accuracy of 77.4%, this means that the model is able to predict 77.4% of the diabetes problem correctly. Also, the model got a high sensitivity of 83 which it is good because it is the ability of a test to accurately classify those with the disease (true positive rate), that is telliing someone they have diabetes disease when they do. While the specificity of 65 is the ability of the test to correctly identify those without the disease (true negative rate), that is telliing someone they do not have diabetes disease when they do not. Although, the specificity is low, this can be improved in future models by increasing the dataset because the data used in training this model was not big enough to draw a general conclusion from.

The table 3 below shows the confusion matrix of the model which was used as a performance metrics to evaluate the model.

Table 3: Confusion matrix

	Reference	
Prediction	No diabetes	Diabetes
No diabetes	127	27
Diabetes	25	52

The table shows that 127 cases do not have diabetes and were predicted correctly (true positive), 27 patients had diabetes but was predicted wrongly as not having diabetes (false

positive), while 25 who did not have diabetes were predicted they did (false negative) and lastly, 52 patients who had diabetes were predicted correctly (true negative).

V. Relevant Features and Future works

During the course of building this predictive model, it was discovered that there are major factors to features to consider when predicting diabetes which will be useful for the model to boost its predictive power.

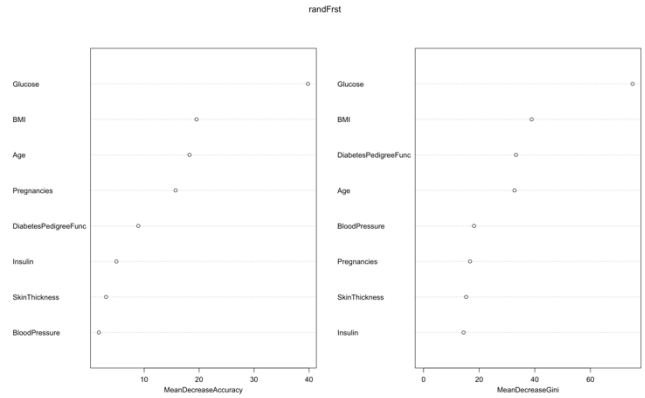

Figure 3: Variables importance

The figure 3 above shows a plot of the important factors in order of importance. using the varImpPlot() function in R, this was possible. From the figure, it is seen that Glucose is the most important factor used in building a diabetes model followed by BMI, Age, Diabetes Pedigree function, Blood pressure and so on. Another feature found was the data sample on diabetes collected contained females between 21 to 81 of which most of them were in their twenties already have diabetes. Diabetes can be prevented if they improve on their diets and perform regular exercise.

The main essence of this research is to explore the predictive analysis. This model can be implemented in the health care system as an automated predictive system for diabetes as this will help Endocrinologist in the healthcare domain in predicting diabetes fast and easy as it requires little or no human intervention. With this model, we can Better care for those with prediabetes (Patient-Centeredness), Refine the treatment pathways for a better-quality approach to medication, Change the care system and Improve financial performance of a nation.

During future models, we seek to utilize a bigger dataset with more features as this plays a major role when building a predictive model. Also, implementing the important variables discovered during this research to boost the accuracy of the model and explore other classification machine learning techniques that can be used in predicting diabetes. One further advancement in this field for future models is to enhance the design of a model for not only predicting diabetes but also differentiating between the different diabetes types with possible treatment or medications. We require thorough research in other to include the treatments after diagnosis in other to automate the treatment systems in the healthcare industry as this will reduce the consultation time and provide efficient results and health in a country as this will reduce diabetes in the nation.

References

[1] N. S. El and S. S. Abu-Naser, 'Diabetes Prediction Using Artificial Neural Network', *Int. J. Adv. Sci. Technol.*, p. 12, 2018.

[2] U. R. Acharya, N. Kannathal, E. Y. K. Ng, L. C. Min, and J. S. Suri, 'Computer-Based Classification of Eye Diseases', in *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*, New York, NY, Aug. 2006, pp. 6121–6124, doi: 10.1109/IEMBS.2006.260211.

[3] H. Kaur and V. Kumari, 'Predictive modelling and analytics for diabetes using a machine learning approach', *Appl. Comput. Inform.*, vol. ahead-of-print, no. ahead-of-print, Jul. 2020, doi: 10.1016/j.aci.2018.12.004.

[4] G. Kaur and A. Chhabra, 'Improved J48 Classification Algorithm for the Prediction of Diabetes', *Int. J. Comput. Appl.*, vol. 98, no. 22, pp. 13–17, Jul. 2014, doi: 10.5120/17314-7433.

[5] A. Mir and S. N. Dhage, 'Diabetes Disease Prediction Using Machine Learning on Big Data of Healthcare', in *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, Pune, India, Aug. 2018, pp. 1–6, doi: 10.1109/ICCUBEA.2018.8697439.

[6] A. Agarwal and A. Saxena, 'Analysis of Machine Learning Algorithms and Obtaining Highest Accuracy for Prediction of Diabetes in Women', p. 5.

[7] V. V. Vijayan and C. Anjali, 'Prediction and diagnosis of diabetes mellitus — A machine learning approach', in *2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS)*, Trivandrum, Dec. 2015, pp. 122–127, doi: 10.1109/RAICS.2015.7488400.

[8] Md. F. Faruque, Asaduzzaman, and I. H. Sarker, 'Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus', in *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, Cox'sBazar, Bangladesh, Feb. 2019, pp. 1–4, doi: 10.1109/ECACE.2019.8679365.

[9] D. Petrovic, R. Roy, and R. Petrovic, 'Modelling and simulation of a supply chain in an uncertain environment', *Eur. J. Oper. Res.*, vol. 109, no. 2, pp. 299–309, Sep. 1998, doi: 10.1016/S0377-2217(98)00058-7.

[10] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, 'Predicting Diabetes Mellitus with Machine Learning Techniques', *Front. Genet.*, vol. 9, p. 515, Nov. 2018, doi: 10.3389/fgene.2018.00515.

[11] M. Khalilia, S. Chakraborty, and M. Popescu, 'Predicting disease risks from highly imbalanced data using random forest', *BMC Med. Inform. Decis. Mak.*, vol. 11, no. 1, p. 51, Dec. 2011, doi: 10.1186/1472-6947-11-51.

[12] W. Xu, J. Zhang, Q. Zhang, and X. Wei, 'Risk prediction of type II diabetes based on random forest model', in *2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)*,

Chennai, India, Feb. 2017, pp. 382–386, doi: 10.1109/AEEICB.2017.7972337.

[13] J. P. Kandhasamy and S. Balamurali, 'Performance Analysis of Classifier Models to Predict Diabetes Mellitus', *Procedia Comput. Sci.*, vol. 47, pp. 45–51, 2015, doi: 10.1016/j.procs.2015.03.182.

[14] N. Nai-arun and R. Moungmai, 'Comparison of Classifiers for the Risk of Diabetes Prediction', *Procedia Comput. Sci.*, vol. 69, pp. 132–142, 2015, doi: 10.1016/j.procs.2015.10.014.

[15] N. Sneha and T. Gangil, 'Analysis of diabetes mellitus for early prediction using optimal features selection', *J. Big Data*, vol. 6, no. 1, p. 13, Dec. 2019, doi: 10.1186/s40537-019-0175-6.