 doi:10.20944/preprints202106.0144.v1

Prediction of Female diabetic Patient in India using different Learning Algorithms

Payal Bose[1][0000-0002-2177-5775], Prof. Samir K Bandyopadhyay[1][0000-0002-4868-3459] and Prof. Vishal Goyal[1]

[1]GLA University, Mathura-Delhi Road Mathura, Chaumuhan, Uttar Pradesh 281406

Abstract: Diabetics or Diabetic Mellitus is a metabolic disorder of blood sugar levels in the human body. It is a major non-communicable disease and involved many serious health risk issues. This disease is rapidly increasing in India. It is a chronic condition and it occurs when a body doesn't produce enough insulin hormone to control the blood sugar level. In this study, different variables have been analyzed that cause the diabetics, and different machine learning algorithms are used to predict whether an unknown sample is diabetes or not. For this purpose, PIMA diabetic detection for Female patients was used. Here 10 different classification model is used for prediction. Finally, the detailed performance analysis of the different variables of the PIMA dataset and also the classification model are discussed.

Keywords: *Diabetic, Diabetic Mellitus, Diabetic Prediction, PIMA diabetic dataset, Female diabetic Patients, Machine Learning*

I. INTRODUCTION

Diabetics Mellitus [1] or simple diabetes is a metabolic disorder or inadequate control of blood sugar levels. In the human body, the hormone insulin helps to move the glucose from blood to the cell and store the glucose as energy for future use. When the percentage of this glucose level is increases or decreases from its normal level, then diabetics are occurring. The high blood sugar level or the high diabetics can damage nerve, eyes, kidneys, skin, and many other organs. This is a very common disease and non-curable. But changing the lifestyle and food habits with treatment can help to maintain the blood sugar level and gives a healthy life. Diabetics come with depending on the different reasons.

a. **Prediabetic.** When the blood sugar level increases from normal to ceratin level but that should not enough to diagnose, then it is called Prediabetic [2]. It causes a heart attack and also the type-2 diabetics. Exercise and body weight maintenance can help to reduce the risk factors in the future.

b. **Type-1 Diabetic.** It is an insulin-dependent diabetic [3]. When the immune system of a body destroys the insulin-making cell then this type of diabetics is occurring. This type of diabetics usually occurs in children and young people. To control this diabetes a certain amount of insulin needs to consume on a daily purpose.

c. **Type-2 Diabetic.** This type of diabetes is insulin-independent [4]. It is commonly observed in adults but for the last 20 years this type of diabetes also common in the young generation due to overweight or obese. This diabetes occurs when an immune system resists to create the insulin hormone and the sugar level in blood is increased abruptly. Maintaining body weight, food habits, daily exercise, and proper medication can help to reduce the risk factors.

d. **Gestational Diabetic.** During pregnancy when some of the hormones resist to produce insulin hormones, then the blood sugar level in the mother's body is increasing. This type of diabetes [5] is often diagnosed during the middle or last stage of pregnancies. This type of diabetes is more riskier for the baby than the mother. Proper medication, enough nutrients can help to reduce the risk factors.

 doi:10.20944/preprints202106.0144.v1

Motivation and Aim of this study. Diabetes is one of the most common and serious health issues today. The occurrence of diabetic patients increasing day by day and most of them are female patients [6]. In several research it was observed that the factors like Body Mass Index, Blood Pressure, Insulin Level, Cholesterol are the main for causing diabetes. For female patients pregnancies is an additional but also an important factor. This study shows the behavior of the different key factors for diabetics patients and also the relationship between the main key factors. The aim of this study is to predict whether a patient is diabetic or not, particularly female patients, based on different machine learning algorithms.

II. LITERATURE REVIEW

In one research article, the authors made a comparative analysis of different machine learning algorithms. They evaluated the performance of the different machine learning algorithms using the PIMA diabetics dataset of female patients in India. They show the random forest classifier gives more than 74% accuracy [7].

The authors in their study used the PIMA diabetics dataset with the other dataset collected from Kurmitola General hospital in Bangladesh. They used different machine learning algorithms to perform the prediction. Finally shows the Naïve Bayes algorithm gives the best result among them [8].

Other authors used this PIMA diabetics dataset in their study to predict the disease. They also used the different machine learning algorithms to calculate the classification accuracy, precision, F1-score, and accuracy under the ROC curves [9].

In one research article, the authors investigate the prediction of diabetics based on the input features FPG and HbA1c. They used five different machine learning algorithms with hierarchical clustering, feature elimination, and feature permutation techniques. They identify different risk factors that are indirectly involved with diabetes classification [10].

The authors in their study used diverse machine learning algorithms on the PIMA diabetic prediction dataset. They used the classifiers Artificial Neural Network (ANN), Naive Bayes (NB), Decision Tree (DT), and Deep Learning (DL) and achieved 90% to 98% accuracy. They also show that the deep learning approach achieved the maximum accuracy, 98.04% [11].

III. DATASET OVERVIEW

a. DATASET

The PIMA diabetic dataset[12] of Indian female patients was downloaded from Kaggle. This dataset was originally collected from the National Institute of Diabetes and Digestive and Kidney Diseases. The purpose of this dataset is to diagnosis whether a patient is diabetic or not based on certain measurement parameters. All the patients in this dataset are female and at least 21 years old of Pima Indian heritage.

b. DATASET DETAILS

This dataset [13] contains a total of 768 data of female patients. Among them, 500 female patients are diagnosed as non-diabetic and 268 female patients are diabetic. The diagnosis result is stored as the binary values in the dataset for each patient with the other attributes. Table 1 and figure 1 show the dataset summary and the diagnosis details of the PIMA dataset.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 June 2021 doi:10.20944/preprints202106.0144.v1

Table 1: PIMA Dataset Summary

PIMA Diabetic Detection dataset	
Total No of Observation	768
Total No of Features	9
Total No of Diabetic Patients	268
Total no of Non-diabetic patients	500
Diabatic : Non-Diabatic (ratio)	0.35 : 0.65

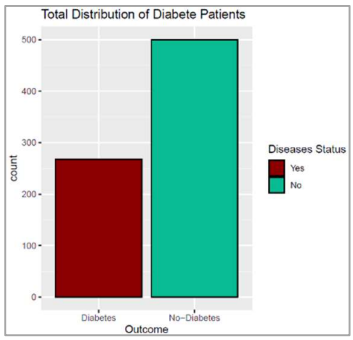

Figure 1:Total no of Diabatic Patient Distribution

c. DATASET FEATURE DETAILS

The dataset has 8 predictive variables and one responsive variable. The variables' details are shown in table 2.

Table 2: Dataset Variables Details

No.	Feature name	Feature Details	Min Value	Max value
1	Pregnancies	Number of times pregnant	0.0	17.0
2	Glucose	Plasma glucose concentration 2 hours in an oral glucose tolerance test	0.0	199.0
3	BloodPressure	Diastolic blood pressure (mm Hg)	0.0	122.0
4	SkinThickness	Triceps skin fold thickness (mm)	0.0	99.0
5	Insulin	2-Hour serum insulin (mu U/ml)	0.0	846.0
6	BMI	Body mass index (weight in kg/(height in m)^2)	0.0	67.1
7	DiabetesPedigreeFunction	Diabetes pedigree function	0.078	2.42
8	Age	In years	21	81
9	Outcome	Class variable (0 or 1)	-	-

d. DATASET ANALYSIS

1. **Pregnancies.** The dataset contains a total of 8 predictive variables. All the patients in this dataset are female so the no of pregnancies is one of the most important variables for analysis. Figure 2a shows the boxplot of no of pregnancies of diabetic and non-diabetic female patients. Figure 3a shows the bar plot of diabetic and non-diabetic patients grouped by the number of time pregnancies of female patients.
2. **Glucose.** Variable glucose is another important predicting variable in this dataset. Here the glucose concentration level is observed after 2 hours of a meal. Generally, the glucose concentration level for a normal person is up to 140 after 2 hours of a meal. The percentage bar plot shows the concentration level of below 140 and above 140 of the female patients. Figure 2b and 3b shows the boxplot and bar plot of glucose concentration level.
3. **Blood Pressure.** Here the blood Pressure level is divided into three categories 1) below 65 mm hg is called low-normal, 2) 65-89 range is called normal and 3) above 89 is higher than normal or Hypertension. Figure 2c shows the boxplot of the blood pressure level of female diabetes and non-diabetes patients and 3c show the percentage of diabetes and non-diabetes patients based on the blood pressure level.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 June 2021 doi:10.20944/preprints202106.0144.v1

4. **Skin Thickness.** Human skin thickness is determined by collagen. It is produced underneath the skin. Depending on the insulin level the skin thickness is determined. Figure 2d shows the boxplot of the female patients and 3d shows the percentage level of skin thickness of the female diabetic and non-diabetic patients.
5. **Insulin.** It is a hormone that keeps balance the blood sugar level in the human body. After 2 hours of consuming a meal, the normal insulin level is less than 200 mu U/ml. Figure 2e shows the box plot of the insulin level of diabetic and non-diabetic female patients and 3e shows the percentage level.
6. **BMI (Body Mass Index).** This variable is used to measure body fat. For a diabetic patient, this is an important measuring parameter. Figures 2f and 3f show the box plot and the percentage level of BMI of the female diabetic and non-diabetic patients.
7. **Diabetic Pedigree Function (DPF).** It is a parameter that gives the report about the patient's family diabetic history. This function provides the relationship between the genetic and non-genetic relatives' diabetic status. Figure 2g and 3g show the box plot and the percentage level of DPF of the female diabetic and non-diabetic patients.
8. **Age.** Age is one of the most important parameters for female patients w.r.t diabetic. Based on the different age groups the value of diabetic predicted parameters is changed. Figure 2h and 3h shows the box plot and the percentage level of the different Age groups of the female diabetic and non-diabetic patients.

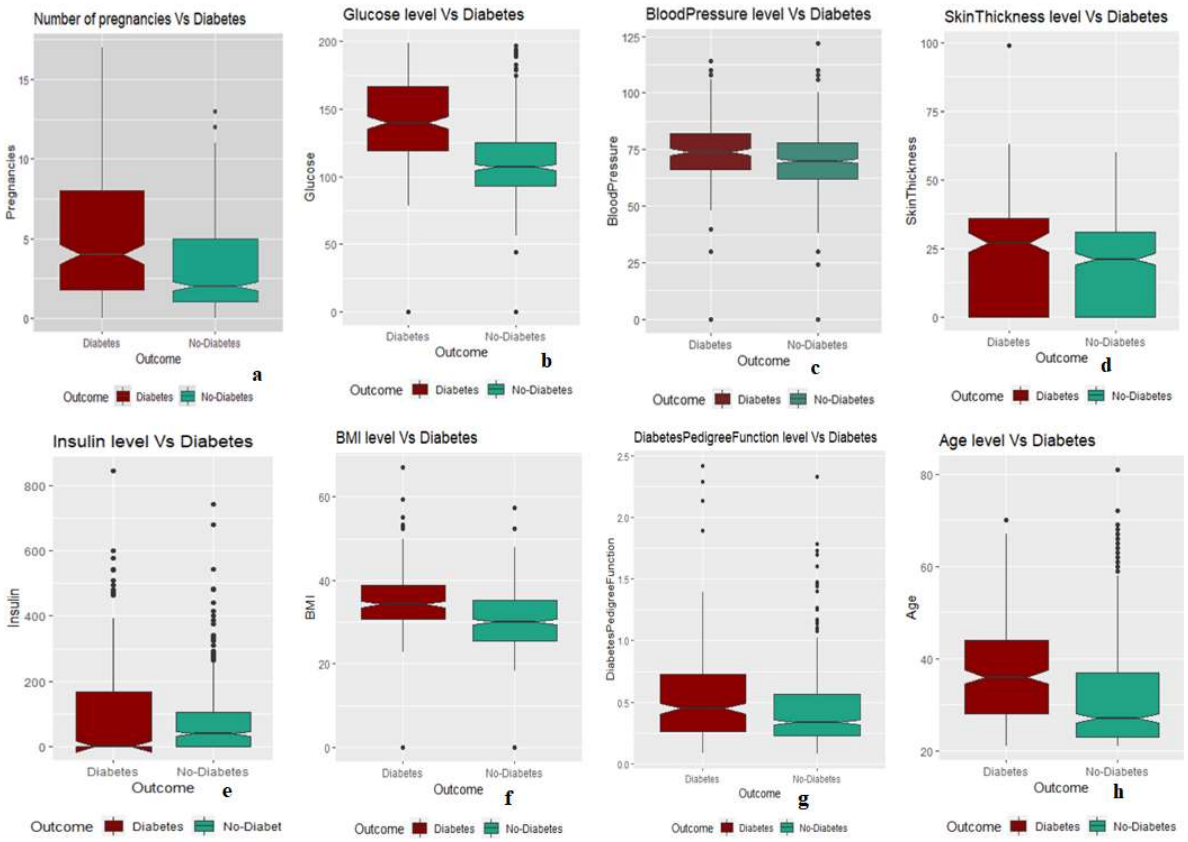

Figure 2 (a-h): Boxplot of different predictive variables of Female Patients

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 June 2021 doi:10.20944/preprints202106.0144.v1

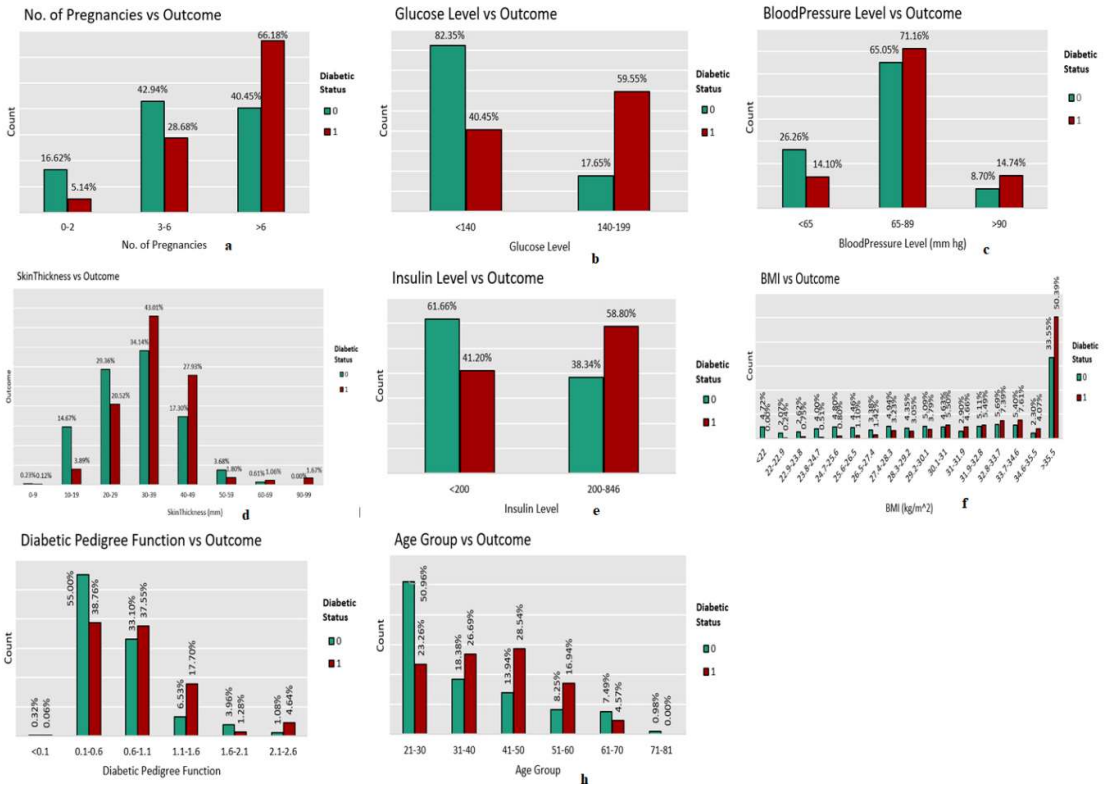

Figure 3 (a-h): Percentage Bar plot of different predictive variables of Female Patients

e. DATASET CORRELATION

A correlation matrix is used to explain the statistical relationship between the variables. Figure 4 shows the correlation of the 8 predictive variables and 1 responsive (Outcome) variable of the PIMA diabetics dataset of female patients in India.

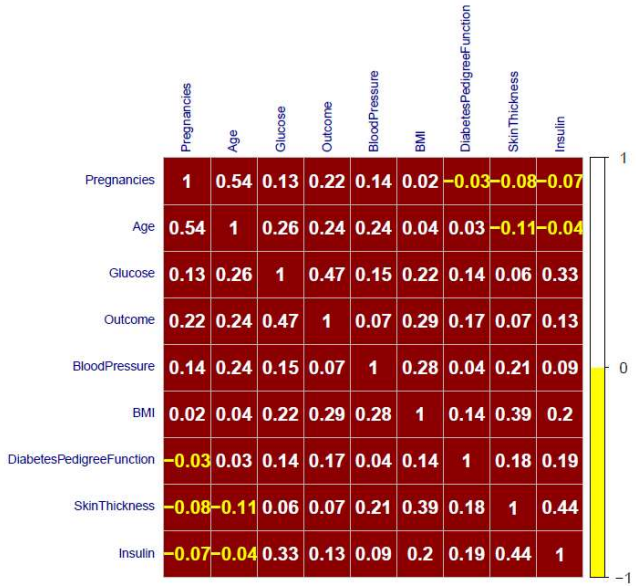

Figure 4: Correlation Plot of PIMA Diabetics Dataset

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 June 2021 doi:10.20944/preprints202106.0144.v1

From the correlation matrix plot, it is observed that the dataset has five important predictive variables. They are, 1) No. of Pregnancies, 2) Glucose level, 3) Insulin level, 4) BMI (Body Mass Index), and 5) Age group. Again, the BMI level also correlated with two other variables 1) Skin Thickness, and 2) Blood Pressure level. The relationship among all the variables is depicted in figure 5 (a-d).

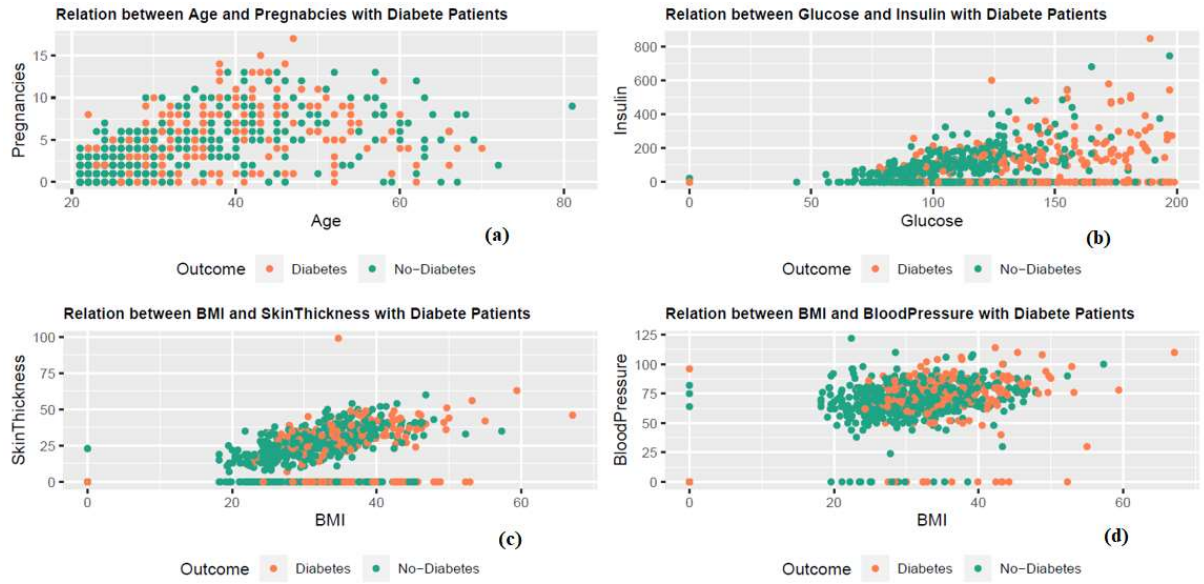

Figure 5(a-d): The relation between the important variables for diabetic prediction

IV. METHODOLOGY

The process to predict a female patient diabetic or not is depicted in figure 6. For this study, 10 different machine learning algorithms are used to check the prediction rate. First, the dataset was pre-processed to be used for the machine learning algorithms. Then the dataset split into training and testing sets. After that, the algorithms are applied to the training set to create the trained dataset. Finally, the trained dataset is applied to the test set to predict the outcome.

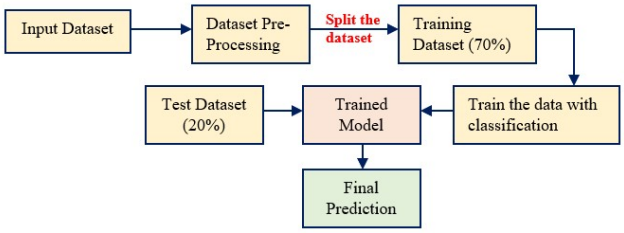

Figure 6: Workflow of the model

1. **Data Pre-processing.** In this part, it is needed to check if any data missing in the original dataset or not. If any data is missing then replaced the null data with the mean values of that variable. After filling in the missing data, the dataset is finally ready for the classification algorithm.
2. **Split the dataset.** After pre-processing the dataset is divided into two parts, 1) Training Dataset to create a Trained Model, 2) Test Dataset. For training the 80% data and for testing 20% data in the whole dataset is used.
3. **Classification Model.** For the prediction of the outcome, in this experiment 10 classification algorithms are used. They are 1) Decision Tree, 2) Random Forest, 3) Gradient Boosting, 4) Logistic Regression, 5) Linear Regression,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 June 2021 doi:10.20944/preprints202106.0144.v1

6) Support Vector Machine, 7) Discriminant Analysis, 8) Quadratic Discriminant Analysis, 9) Mixed Discriminant Analysis and 10) K-Nearest Neighbour Algorithm.

V. EXPERIMENTAL RESULT

All classifications result shown in table 3. The performance of all the classifiers was analyzed based on the training, testing accuracy. All the accuracy is calculated using the below formula.

$$Accuracy = \frac{T_P + T_N}{T_P + F_P + T_N + F_N} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

Where, T_P = (True positive) = Actual Positive Class and Positive Prediction, T_N = (True Negative) = Actual Positive Class and Negative Prediction, F_P = (False Positive) = Actual Negative Class and Positive Prediction, F_N = (False Negative) = Actual Negative Class and Negative Prediction.

Table 3: Performance Analysis of all the Classifiers

Methods	Training Accuracy	Testing Accuracy
Decision Tree	0.8093645	0.7470588
Random Forest	1.0000000	0.8176471
Gradient Boosting	0.8227425	0.7117447
Logistic Regression	0.7909699	0.7294118
Linear Regression	0.7742475	0.7176471
Support Vector Machine	0.7575251	0.8352941
Discriminant Analysis	0.7491639	0.8000000
Quadratic Discriminant Analysis	0.7419963	0.8000000
Mixed Discriminant Analysis	0.7591973	0.8176471
K-Nearest Neighbour	0.7956429	0.8000000

VI. CONCLUSION

In this study, PIMA diabetic detection dataset was used to predict diabetics in female patients. A detailed analysis of 8 predictive variables and one responsive variable was done in this study. After that, a correlation matrix was calculated to show the important variables and their dependencies variable. Various supervised machine learning algorithms are used to calculate the prediction performance. Finally, a detailed comparative analysis of the algorithms is also performed. Here only one dataset is used and the number of observations in this dataset is quite small. Therefore, the future work is to use a large dataset and also apply the deep learning algorithms for better performance.

ACKNOWLEDGMENTS

Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In Proceedings of the Symposium on Computer Applications and Medical Care (pp. 261--265). IEEE Computer Society Press.

REFERENCES

[1] MedicalNewsToday. What to Know about PCOS and Diabetes. Available online: https://www.medicalnewstoday.com /articles/326185 (accessed on 29 August 2020).

[2] Watson, Stephanie. "Diabetes: Symptoms, Causes, Treatment, Prevention, and More." Healthline, Healthline Media, 27 Feb. 2020, www.healthline.com/health/diabetes.

[3] IDF. Type 1 Diabetes. Available online: https://www.idf.org/aboutdiabetes/type-1-diabetes.html (accessed on 20 January 2020).

[4] IDF Diabetes Atlas, A.D. Type 2 Diabetes. Available online: https://www.idf.org/aboutdiabetes/type-2-diabetes.html (accessed on 20 March 2020).

[5] IDF. Gestational Diabetes. Available online: https://www.idf.org/our-activities/care-prevention/gdm.html (accessed on 1 June 2020).

[6] WHO. Diabetes. Available online: https://www.who.int/news-room/fact-sheets/detail/diabetes (accessed on 1 June 2020).

[7] Varma, K.M.; Panda, D.B.S. Comparative analysis of Predicting Diabetes Using Machine Learning Techniques. J. Emerg. Technol. Innov. Res. **2019**, 6, 522–530.

[8] Pranto, B., Mehnaz, S. M., Mahid, E. B., Sadman, I. M., Rahman, A., & Momen, S. (2020). Evaluating machine learning methods for predicting diabetes among female patients in Bangladesh. *Information (Switzerland)*, *11*(8). https://doi.org/10.3390/INFO11080374

[9] Kumar, S., Kumar, S., Kumar, K., & Abhisekh, P. A. (2021). *Prediction of Diabetes in Females of Pima Indian Heritage : A Complete Supervised Learning Approach*. *12*(10), 3074–3084.

[10] Ahmad, H. F., Mukhtar, H., Alaqail, H., Seliaman, M., & Alhumam, A. (2021). Investigating health-related features and their impact on the prediction of diabetes using machine learning. *Applied Sciences (Switzerland)*, *11*(3), 1–18. https://doi.org/10.3390/app11031173

[11] Naz, H., & Ahuja, S. (2020). Deep learning approach for diabetes prediction using PIMA Indian dataset. *Journal of Diabetes and Metabolic Disorders*, *19*(1), 391–403. https://doi.org/10.1007/s40200-020-00520-5

[12] PIMA. University of California, Irvine Learning Repository. Available online: https://www.kaggle.com/uciml/pima-indians-diabetes-database (accessed on 6 October 2020).

[13] Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In Proceedings of the Symposium on Computer Applications and Medical Care (pp. 261--265). IEEE Computer Society Press.