Original Research

Volume 5, Issue 1, pp 1-9, October, 2023

ACHIEVERS JOURNAL OF SCIENTIFIC RESEARCH

Open Access Publications of Achievers University, Owo

Available Online at *www.achieversjournalofscience.org*

Prediction of Diabetes Using Machine Learning Techniques

AKINBOHUN, F.[1] and ADEGUN, I.P.[2,*]

[1] Department of Computer Science, Rufus Giwa Polytechnic, Owo, Nigeria
[2]Department of Computer Science, Federal University of Technology Akure, Nigeria
Corresponding author e-mail: iyanupelumi22@gmail.com, ipadegun@futa.edu.ng

Submitted: May 12, 2023; Revised: September 22, 2023; Accepted: October 17, 2023; Published: October 20, 2023

Abstract

Diabetes Mellitus is a metabolic disorder that occurs when the body cannot produce sufficient insulin. Its prevalence has seen a significant surge worldwide, necessitating improved methods for early and accurate prediction. Machine learning techniques have proven to be effective in the prediction of diabetes. This study harnesses the capabilities of machine learning (ML) techniques to predict diabetes. To improve the learning efficiency and prediction performance, feature selection techniques were employed in the study. This process selects only optimal features that contributes the most to prediction variables from entire feature set. In this study, three machine learning algorithms (Support Vector Machine, random forest and decision tree) were applied on Pima Indians diabetes dataset. Consistency and correlation-based feature selection techniques were applied on the dataset to improve prediction performance and reduce dimensionality. The results from the experiments show that of all the three models that were used, there was a significant improvement in the performance of the models when feature selection techniques were used. For instance, Support Vector Machine had an accuracy of 81.74% before feature selection as opposed to the accuracy of 79.13% before its application. Random Forest also had an accuracy of 80.08% using Consistency feature selection method as opposed to an accuracy of 77.78% before its application.

Keywords: Diabetes Mellitus, Consistency-based Feature Selection, Correlation-based Feature Selection, Machine learning, Prediction.

1.0 Introduction

Diabetes Mellitus is one of the chronic ailments that many people all over the world suffer from Diabetes often leads to other medical conditions such as heart problems, kidney problems, stroke, hypertension, cardiovascular dysfunction, retinal failure, cerebral vascular dysfunction or even death (Hasan *et al.*, 2017; Mujumdar and Vaidehi, 2019). Statistical reports from the World Health Organization (WHO) revealed that a total of 422 million people in the world are living with diabetes (Mujumdar and Vaidehi, 2019). Another report showed that about 1.5 million people lose their lives because of diabetes (WHO, 2022). Diabetes has been discovered to have only a short-term cure, however, it can be prevented and controlled when detected early enough (Hasan *et al.*, 2017). Findings made by the Center for Disease Control and Prevention (CDC) showed that, medically, diabetes can be detected through conducting laboratory tests like fasting blood sugar test, glucose screening test and random blood sugar tests (CDC, 2022). These medical tests can be time consuming, error- prone and difficult.

In recent times, machine learning has helped in automation of many processes, thereby eliminating human efforts. Machine learning involves the use of certain computer algorithms to make predictions by learning from data rather than through explicit programming. A machine learning model is the output generated when machine learning algorithm is trained with data (Dutta *et al.*, 2018). Examples of existing machine learning algorithms that have been used in predicting diabetes include Decision Tree, Random Forest, Artificial Neural Network, Support Vector Machine and many others. Machine learning techniques have proven to be effective in the prediction of diabetes. Many studies have been conducted on diabetes prediction using machine learning techniques (Sawar *et al.*, 2018).Sonar and Jayamlini (2019) conducted an analysis to determine the risk level of a patient with diabetes. They used artificial neural network, naïve bayes, decision tree and support vector machine for their experiment. Their result showed that the models performed well in determining the risk level of diabetic patients. Khanam and Foo (2021) also tested seven different machine learning techniques for prediction of diabetes using Pima Indian dataset. Their result showed that logistic regression and support vector machine had the best prediction performance.

One major challenge is that several existing diabetes datasets often contains irrelevant and redundant features which reduces prediction accuracy. Feature selection is one of the pre-processing technique that involves identifies the most relevant features in a given dataset. Since only optimal features contribute the most to prediction variables, prediction performance is improved and dimensionality (Remeseiro and Bolon-Canedo, 2019). Hence, this study seeks to apply feature selection to diabetes prediction process. This will help in reducing gover fitting of data and training time, thereby leading to an overall improvement in prediction accuracy. Three machine learning algorithms were used to build our predictive models: Support Vector Machine (SVM), decision tree and random forest algorithm.

2.0 Methodology

To develop the predictive model for Diabetes prediction, four major phases were involved. These include data collection, data pre-processing (feature selection), model construction, model evaluation and prediction as shown in Figure 1.

2.1 Data Collection

Dataset used for this study was acquired from the Pima Indian diabetes dataset available at UCI machine learning repository downloaded from Kaggle website (UCI Machine Learning, 2016). The dataset helps in predicting whether a patient has diabetes based on certain diagnostic parameters within the dataset. The dataset contains 768 records or instances and 8 predictors and 1 class or target. Table 1 shows a brief description of the dataset of the Diabetes mellitus which contains the attributes and their descriptions. This dataset was divided into training data and testing data. 70% of the instances were used for training while the remaining 30% of the instances were used for testing.

2.2 Feature Selection phase

Feature selection plays a very essential role in machine learning tasks. The purpose of using feature selection methods is to remove redundant and irrelevant features from the dataset. In this study, two feature selection methods, namely Correlation Feature Selection and Consistency Based Feature Selection Methods, were deployed.

2.3 Models Construction phase

After feature selection, the next phase in the proposed system for predictive model of Diabetes Mellitus is the building the model. This study used three classification algorithms namely Decision Tree (C45), Random Forest and Support Vector Machine. The models are constructed as follows:

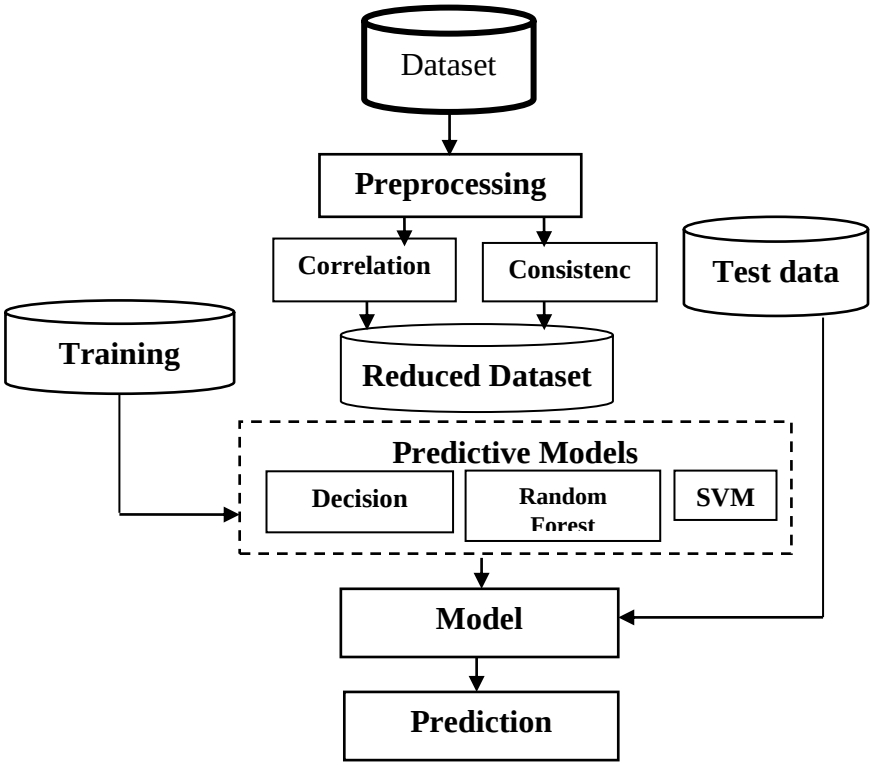

Figure 1: System Architecture

Table 1: Description of the Pima Indian Dataset

No	Attribute	Description
1	Pregnancies (NoP)	Number of times pregnant
2	Glucose (PLC)	Plasma glucose concentration 2 hours in an oral glucose tolerance test
3	Blood Pressure (DBP)	Diastolic blood pressure(mmHg)
4	Skin Thickness (TST)	Skin Thickness
5	Insulin (HAI)	2-Hour serum insulin (mu U/ml)
6	BMI	Body Mass Index (BMI)
7	Diabetes Pedigree Function (SPF)	Diabetes Pedigree function
8	Age (Age)	Age (in years)
	Class	0(No) or 1(Yes)

Decision Trees (C4.5): Decision Tree was used to determine whether a patient has diabetes or not for a given instance. The training data were assumed to be represented as a pair [x_1, x_2, x_3, …x_n→ y] where x_1, x_2, x_3 …x_n are the predictor describing some instances while y is the appropriate class or target.

The basic principle behind Decision Tree (c45) is described as follows. C45 uses Information Gain where entropy for each branch is calculated (i.e., the entropy of the class and each subset of the attribute/feature) are computed using the equation (1)

$$E = -\sum_{i=1}^{n} P_i log2P_i \quad (1)$$

where $P_i$represents the proportion of instances in the Pima Indian dataset which belong to the $i-th$ class, and n representsthe number of classes.

Decision Tree (C4.5) uses Information Gain Ratio whose attributes are computed using Equation (2). The attribute with the maximum gain ratio value is finally chosen as the splitting attribute.

$$\text{Gain Ratio}_{\text{attribute}} = \frac{\text{Gain}_{\text{attribute}}}{\text{SplitInfo}_{\text{attribute}}} \quad (2)$$

Random Forest: Is an ensemble supervised learning algorithm that can perform both regression and classification tasks. Random forest algorithms are based on decision trees which uses several "if…. then" branches before arriving at its final decision (branch). Other features of random forest includes dimensional reduction, its ability to fix missing and to handle outliers.

Support Vector Machine: Support Vector Machine is a supervised machine learning algorithm that is often used for binary classification. It can also be used to perform regression tasks. SVM uses a hyperplane to separate the data into their appropriate classes. In situations where there are multiple hyperplanes, the one with the biggest margin is selected as the most correctly classified. The hyperplane is a function like the equation for a line. SVM is expressed mathematically as follows:

Given a two-class problem represented as $\{x_i, y_i\}_{i=1}^{N}$, where x_i and y_i represents input and output vectors respectively and $y_i \in \{-1, +1\}$ is the class label of input x_i. SVM seeks to find an optimal decision boundary (hyperplane) that classifies all data points correctly. The hyperplane is expressed in equation (3).

$$w(x_i) + b = 0 \quad (3)$$

where w stands for optimal set of weights and b represents the optimal bias.

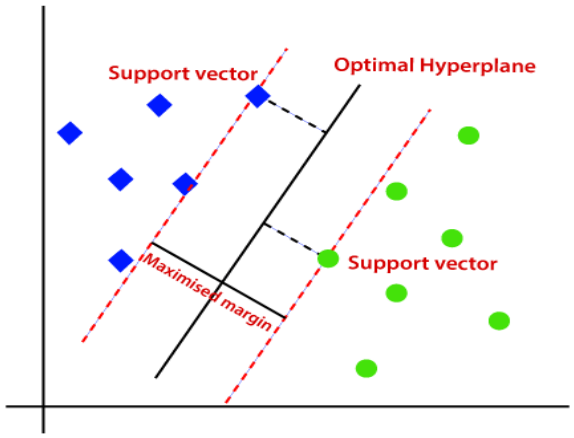

Figure 2: Description of Support Vector Machine (Adopted from Javatpoint, 2022)

Model Evaluation: The models were evaluated using Accuracy, Precision, Recall and F1 Score as performance metrics. We based the performance evaluation on F1 Score because it (F1 Score) finds the harmonic mean of Precision and Recall.

3.0 Results and Discussion

The study was setup in two stages: The first stage was that the dataset was trained without applying the two feature selection methods on the Pima Indian Dataset. The second stage involved application of consistency-based and correlation-based feature selection methods to the same dataset.

3.1 Evaluation Results without Feature Selection

The dataset was trained and tested using decision tree, random forest and SVM algorithms without applying feature selection methods. The result shows that Support Vector Machine had accuracy of 79.19% followed by Random Forest with accuracy of 77.78% and Decision Tree had the least accuracy of 76.52% as shown in Table 2 and Figure 3. Considering the F1 Score metric, being a good performance metric; SVM had highest value of 0.782 followed by Random Forest with F1 Score of 0.772 and Decision Tree had the lowest F1 Score of 0.771.

Table 2: Performance metrics on the Dataset without Feature Selection Methods**.**

Performance Metric	Decision Tree (C45)	Random Forest	SVM
Accuracy (%)	76.52	77.78	79.13
Precision	0.79	0.77	0.78
Recall	0.77	0.77	0.79
F1 Score	0.77	0.77	0.78

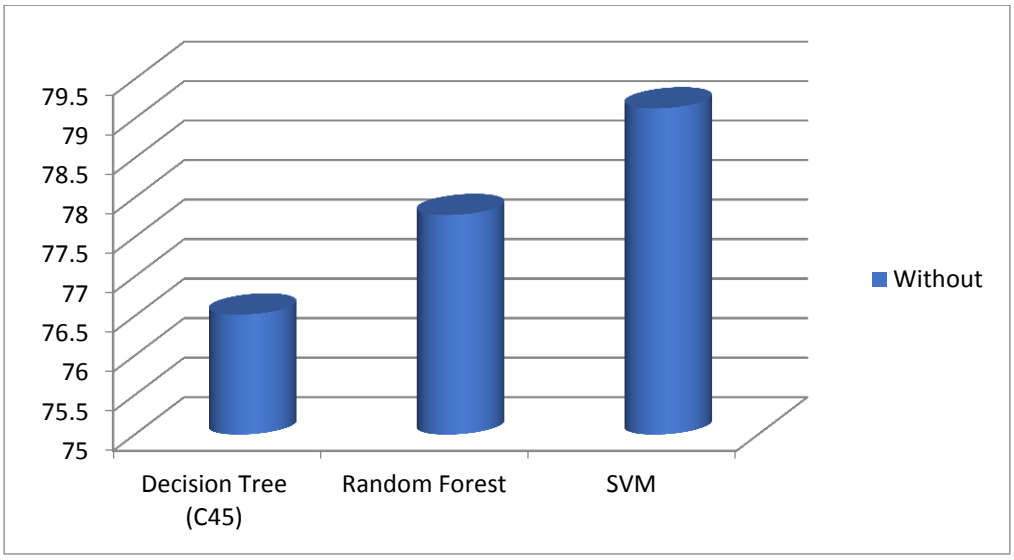

Figure 3: Accuracy of the classifiers without Feature Selection

3.2 Evaluation Results with consistency-Based Feature Selection

The selected features using Consistency-based feature selection are pregnancies, glucose, blood pressure, diabetes pedigree function. The results of the three models namely Decision Tree (C45) Random Forest and Support Vector Machine on the reduced datasets are presented in Table 3 and 4. Table 3 shows the result when Consistency method was used on Decision Tree (C4.5), Random Forest and Support Vector Machine. The application of Consistency method on the three models showed that SVM had accuracy of 81.74% followed by Random Forest with accuracy of 77.39% and Decision Tree had the least value of 76.96%. The F1 Score of the three

models: SVM, Random Forest and Decision Tree were: 0.812, 0.772 and 0.77 respectively. SVM had the highest F1 Score.

Table 3: Evaluation Result with Consistency Feature Selection Method

Performance metric	**Decision Tree (C45)**	**Random Forest**	**SVM**
Accuracy (%)	76.96	77.39	81.74
Precision	0.77	0.77	0.81
Recall	0.77	0.77	0.81
F1 Score	0.77	0.77	0.81

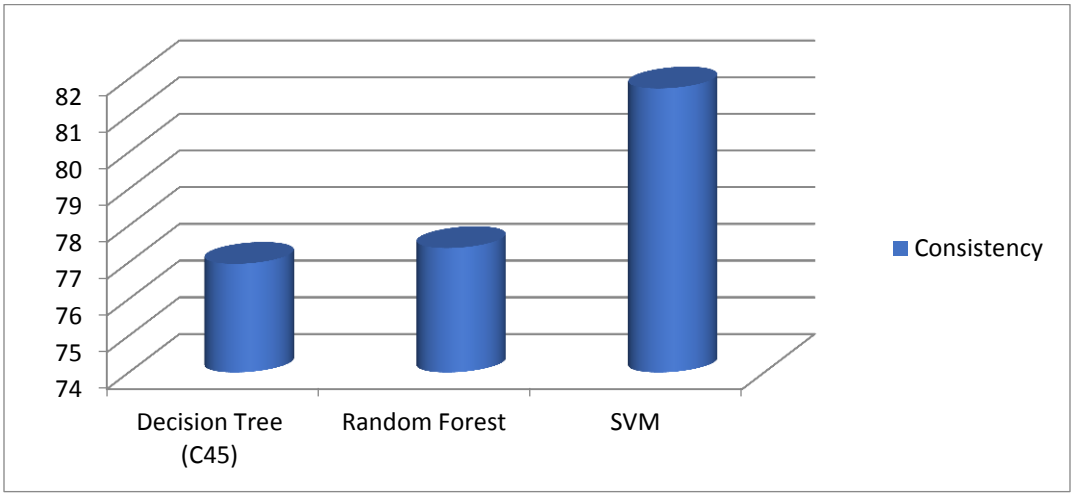

Figure 4: Accuracy of the classifiers with Consistency Feature Selection

3.3 Evaluation Results with Correlation-Based Feature Selection

When Correlation model was deployed as Feature Selection method, the following attributes were selected: Glucose, Body Mass Index (BMI), Age, Number of Pregnancies, Diabetes Pedigree Function and Insulin. Table 4 shows the results of using Correlation method on the Decision Tree (C4.5), Random Forest and Support Vector Machine.

When the dataset was trained and tested with Correlation method on the models namely: The Random Forest had the highest accuracy of 80.08%, followed by SVM with accuracy of 79.13% and the least was Decision Tree of accuracy 77.78%. The F1 Scores of Random Forest, Decision Tree and SVM were 0.796, 0.783 and 0.782 respectively. Hence Random Forest had the highest F1 score while SVM and Decision Tree were close in value by 0.001.

Table 4: Evaluation Result with Correlation Feature Selection Method

Performance metric	**Decision Tree (C45)**	**Random Forest**	**SVM**
Accuracy (%)	77.78	80.08	79.13
Precision	0.80	0.79	0.78
Recall	0.78	0.80	0.79
F1 Score	0.78	0.80	0.78

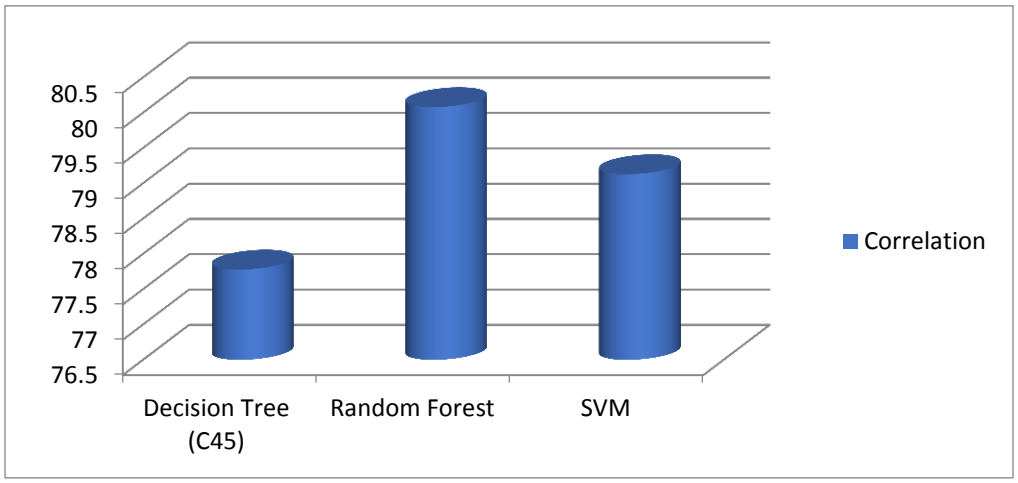

Figure 5: Accuracy of the classifiers with Correlation Feature Selection

3.3 Comparative Analysis of the Three Models with Feature Selection and Without Feature Selection

The result of models after applying the two feature selection methods was compared to know which was better in predicting diabetes. Out of the two feature selection methods used, it was observed that Consistency method performed better than Correlation method in Diabetes Mellitus prediction. SVM performed best among the three models. Overall, there was an improvement in the performance of the models (SVM, Random Forest and Decision Tree) when the two feature selection techniques were applied as shown in Figure 6. Considering all the evaluation results, Consistency feature selection method and SVM classifier produces the best prediction results.

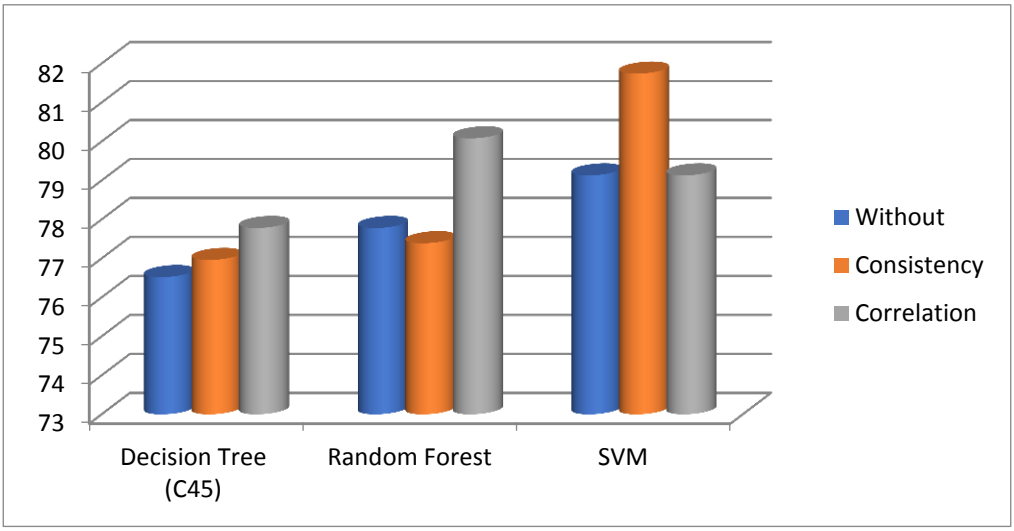

Figure 6: Comparative result of the models before applying feature selection and after applying Feature Selection

4.0 Conclusion and Recommendation

In this study, three machine learning algorithms (Support Vector Machine, random forest and decision tree) were applied on Pima Indians diabetes dataset. Consistency and correlation-based feature selection techniques were applied on the dataset to improve prediction performance and reduce dimensionality. The results from the experiments show that of all the three models that were used, support vector machine had the highest prediction performance before feature selection was applied and after its application. This shows that SVM is an effective algorithm for diabetes prediction. The result also showed some improvement when feature selection was applied. Further studies can investigate the use of ensemble of different machine learning techniques and the use of some other feature selection techniques.

References

Center for Disease Control and Prevention (2022). National Diabetes Statistics Report. Retrieved from https://www.cdc.gov/diabetes/data/statistics-report/index.html on 19th January 2023.

Dutta, D., Paul, D. and Ghosh, P. (2018). Analysing feature importances for diabetes prediction using machine learning. In *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)* (pp. 924-928). IEEE.

Hasan, M. K., Alam, M. A., Das, D., Hossain, E., and Hasan, M. (2020). Diabetes prediction using ensembling of different machine learning classifiers. *IEEE Access*, *8*,76516-76531.

Khanam, J.J. and Foo, S.Y. (2021). A comparison of machine learning algorithms for diabetes prediction. *ICT Express*, *7*(4), 432-439.

Mujumdar, A. and Vaidehi, V. (2019). Diabetes prediction using machine learning algorithms. *Procedia Computer Science*, *165*, 292-299.

Remeseiro, B. and Bolon-Canedo, V. (2019). A review of feature selection methods in medical applications. *Computers in biology and medicine*, *112*, 103375.

Sarwar, M.A., Kamal, N., Hamid, W. and Shah, M.A. (2018). Prediction of diabetes using machine learning algorithms in healthcare. In *2018 24th international*

conference on automation and computing (ICAC) (pp. 1-6). IEEE.

Sonar, P. and JayaMalini, K. (2019). Diabetes prediction using different machine learning approaches. In *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 367-371). IEEE.

UCI Machine Learning. (2016). Pima Indians diabetes database. *kaggle. com/uciml/pima-indians-diabetes database*.

World Health Organization Newsroom (2022). Diabetes Key Facts. Retrieved from https://www.who.int/news-room/fact-sheets/detail/diabetes. Accessed 19th January 2023.