Research on Diabetes Prediction Model of Pima Indian Females

Yining Guan
Eberly College of Science,
Pennsylvania State University,
University Park, PA 16802, USA
ymg5169@psu.edu

Chia Jung Tsai
Susan and Henry Samueli College of
Health Sciences, University of
California Irvine, Irvine, CA
92697-3950
chiajt@uci.edu

Shuyuan Zhang[*]
Centre for Misfolding Diseases, Yusuf
Hamied Department of Chemistry,
University of Cambridge, Cambridge,
CB2 1EW, UK
1510493@mail.nankai.edu.cn

ABSTRACT

Diabetes mellitus is a widespread global health issue with significant consequences and economic burdens. The Pima Indians, an indigenous community residing in certain regions, have faced significant challenges in accessing adequate healthcare due to poor resources, low income, and unfavorable economic conditions. Consequently, these barriers have led to delayed detection and management of chronic illnesses, particularly diabetes. This study aims to create a diabetes prediction model for Pima Indian females using machine learning algorithms. The dataset includes 768 Pima Indian female participants, with 268 diagnosed with diabetes and 500 without diabetes. After cleaning the data and selecting relevant variables, binomial logistic regression and regression trees were used for modeling. The results revealed that both logistic regression model and regression tree model have the similar accuracy which logistic regression showed better accuracy performance, with an accuracy of 77.48% through cross-validation. Key health indicators like age, body mass index (BMI), glucose level, and diabetes pedigree function were considered as predictors. And from the regression tree, glucose level and BMI might be more strongly associated with the likelihood of having diabetes outcome. By utilizing machine learning algorithms models, members of the Pima community can effectively gauge their risk factors for diabetes and other health conditions, empowering them to take proactive measures for self-monitoring and timely medical interventions when necessary. This research might potentially contribute to simple and low-cost self-prediction of diabetes among Pima Indian females, enabling establishing proactive prevention strategies and interventions by easily accessible indicators to improve health outcomes for at-risk individuals.

CCS CONCEPTS

• **Applied computing** → Computing methodologies; Mathematics of computing.

KEYWORDS

Diabetes, Pima Indian females, Logistic regression, Regression tree

ACM Reference Format:
Yining Guan, Chia Jung Tsai, and Shuyuan Zhang. 2023. Research on Diabetes Prediction Model of Pima Indian Females. In *2023 4th International Symposium on Artificial Intelligence for Medicine Science (ISAIMS 2023), October 20–22, 2023, Chengdu, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3644116.3644168

1 INTRODUCTION

Diabetes, also known as diabetes mellitus, is a chronic metabolic disease that affects millions of people worldwide. It is a complex and multifactorial condition characterized by high blood glucose levels brought on by deficiencies in insulin secretion, action, or both. Diabetes poses significant challenges to public health due to its prevalence, associated complications, and economic burden [1]. Globally, the prevalence of diabetes has risen to epidemic levels, with 463 million adults estimated to have the disease in 2019 [2]. If practical preventive measures are not taken, this number is predicted to increase to 700 million by 2045 [3]. The burden of diabetes extends beyond individuals, as it impacts families, communities, and healthcare systems. As a result of its complications, the quality of life is significantly decreased, and the rates of morbidity and mortality are also raised.

Understanding the underlying causes and risk factors of diabetes is crucial for effective prevention and management. While genetic predisposition plays a role, lifestyle factors, such as sedentary behavior, unhealthy diets, and obesity, are major contributors to the development of type 2 diabetes, which accounts for many diabetes cases worldwide [4]. Diabetes can also be more likely to develop in people who have gestational diabetes, impaired glucose tolerance, and other medical conditions [5]. The pancreatic hormone insulin interacts intricately with a number of target tissues, especially adipose tissue, the liver, and skeletal muscle, as part of the pathophysiology of diabetes. Diabetes disrupts the normal insulin response, which results in abnormal glucose metabolism and high blood sugar levels.

Pima females have long been recognized as a population with a disproportionately high risk of developing diabetes mellitus [6]. The Pima people, a Native American tribe residing primarily in the southwestern United States, have faced a unique set of challenges throughout their history that have contributed to this increased susceptibility. Understanding the interaction of genetics, lifestyle, and environmental factors in the onset of this complex disease is made possible by research on Pima females and their elevated risk of diabetes. Historically, the Pima people adapted to survive in an environment characterized by limited resources and periods

[*] All the authors contributed equally and their names were listed in alphabetical order.

ISAIMS 2023, October 20–22, 2023, Chengdu, China

ACM ISBN 979-8-4007-0813-8/23/10
https://doi.org/10.1145/3644116.3644168

of food scarcity. This adaptive response manifested as a thrifty metabolism, allowing their bodies to efficiently store and utilize energy during times of abundance. However, in today's modern society, where food is readily accessible and sedentary lifestyles prevail, this thrifty metabolism has become a double-edged sword [7]. Particularly among Pima women, it has been noted that a high prevalence of obesity and insulin resistance—both major risk factors for the onset of diabetes—is present.

The increased susceptibility of Pima females to diabetes is largely due to genetic factors. According to studies, the Pima population carries particular gene variants that are linked to decreased insulin sensitivity and impaired insulin secretion, which increases their risk [8]. The intergenerational transmission of these genetic factors, combined with the adoption of modern lifestyles characterized by poor dietary choices and limited physical activity, creates a perfect storm for the development of diabetes [9]. The consequences of diabetes in Pima females are far-reaching and pose significant health challenges. Cardiovascular diseases, kidney complications, retinopathy, and neuropathy are just a few of the devastating complications that can arise from uncontrolled diabetes [10]. Additionally, gestational diabetes, high birth weight, preterm birth, and preeclampsia are all associated with an increased risk of negative outcomes during pregnancy in Pima females with diabetes [11].

Addressing the high risk of diabetes in Pima females requires a comprehensive and multidimensional approach. Prevention efforts should focus on promoting healthy lifestyle choices, including regular physical activity and balanced nutrition, to combat obesity and improve insulin sensitivity. Interventions that are sensitive to culture and honor the traditions and heritage of the Pima people can encourage community involvement and a sense of ownership in the prevention and treatment of diabetes [12]. Additionally, early screening and diagnosis are crucial to identifying individuals at risk and implementing timely interventions. Regular medical check-ups, including monitoring of glucose levels and other relevant biomarkers, can aid in the early detection of prediabetes and diabetes, enabling the initiation of appropriate interventions to prevent or delay the onset of complications [13]. In order to train models and assess the relationship between various risk factors and the likelihood of developing diabetes in the Pima female population, this research project will make use of machine learning algorithms. Through comprehensive analysis of patterns and trends within the data, our models will generate accurate predictions and valuable insights, benefiting both healthcare professionals and individuals. The ultimate objective of this project is to facilitate early detection and proactive management of diabetes among Pima females, enabling healthcare providers to implement targeted prevention strategies and interventions that enhance health outcomes for at-risk individuals.

2 METHODS

2.1 Data Sources and Description

The main data of this study is the data set of diabetes health indicators from Kaggle official website, which originally comes from the National Institute of Diabetes and Digestive and Kidney Diseases. It includes 768 samples that are females at least 21 years old of Pima Indian heritage, which contains 500 Pima Indian female samples that did not suffer from diabetes (represented as "0" in the Outcome), and 268 samples that are diagnosed with diabetes (represented as "1" in the Outcome). The predictor variables include certain diagnostic measurements, which are the number of pregnancies the patient has had, their BMI (normal: below 25kg/m^2; overweight: between 25 and <30 kg/m^2; obesity: >30 kg/m^2), insulin level, glucose level, blood pressure level, skin thickness level, diabetes pedigree function, and age. All of these predictors are quantitative variables, measured when determining a diabetes diagnosis outcome.

2.2 Variable Description

Several factors are closely related to the development and management of diabetes. First, the number of pregnancies a person has had could impact their diabetes risk, multiple pregnancies and a history of gestational diabetes increases the likelihood of developing type 2 diabetes. Glucose level, or blood sugar level, plays a critical role in diabetes as elevated levels over time could lead to complications and organ damage. Insulin, a hormone involved in glucose metabolism, is central to diabetes, with type 1 diabetes characterized by an absolute insulin deficiency and type 2 diabetes associated with insulin resistance or inadequate insulin production. Body Mass Index (BMI), a measure of body fat, is strongly linked to diabetes, as obesity contributes to insulin resistance and impaired glucose metabolism. The diabetes pedigree function evaluates family history to assess genetic predisposition, as having close relatives with diabetes increases an individual's risk. Age is also an important factor, with the prevalence of diabetes increasing with advancing age due to age-related physiological changes and lifestyle factors. Understanding and considering these factors are crucial for identifying diabetes risk, implementing preventive measures, and managing the disease effectively. The specific variable information is shown in Table 1.

2.3 Data Cleaning and Variable Selection

Since there are some unreasonable zero values that appear in the glucose, blood pressure, skin thickness, Insulin, and BMI variables, they are considered as missing values and being filtered out. The observations decline from 768 to 392 as Figure 1 shown.

Moreover, backward selection was developed to determine which variables are the most relevant features to fit a model. According to the result, the variables skin thickness, blood pressure, and insulin are filtered out. The final AIC value is 356.9, which demonstrates this selection is efficient.

2.4 Research Methods

Two methods used in the following data processing and modeling in this study are binomial logistic regression and regression trees, which will be introduced in this section. Among them, binomial logistic regression can help to fit a model and predict whether an instance belongs to diabetes or non-diabetes groups by using other health factors, and regression trees can help to improve predictive accuracy and handle complex relationships in the data.

2.4.1 Binomial Logistic Regression. Binomial regression is a statistical technique used to analyze and model the relationship between one or more independent variables and a binary or categorical

Table 1: Variable description information of Pima Indian Dataset

	Description	Unit	Range	Collection method
Pregnancies	Number of times pregnant	Times	0-17	Demographic
Glucose	Plasma glucose concentration at 2 hours during an oral glucose tolerance test.	Mmol/L	0-199	Laboratory Test
Blood Pressure	Diastolic Blood Pressure	mmHg	0-122	Laboratory Test
Insulin	2 Hour serum insulin	Mu/ml	0-846	Laboratory Test
BMI	Body Mass index	(Kg/m^2)	0-67.1	Demographic
Diabetes Pedigree	Diabetes risk function that utilizes family history of diabetes to derive an individual's risk value for developing diabetes.	Percentage	0.08-2.42	Demographic
Age	Age	Years old	21-81	Demographic
Outcome	Diabetes	1 Yes 0 No	0/1	Laboratory Test

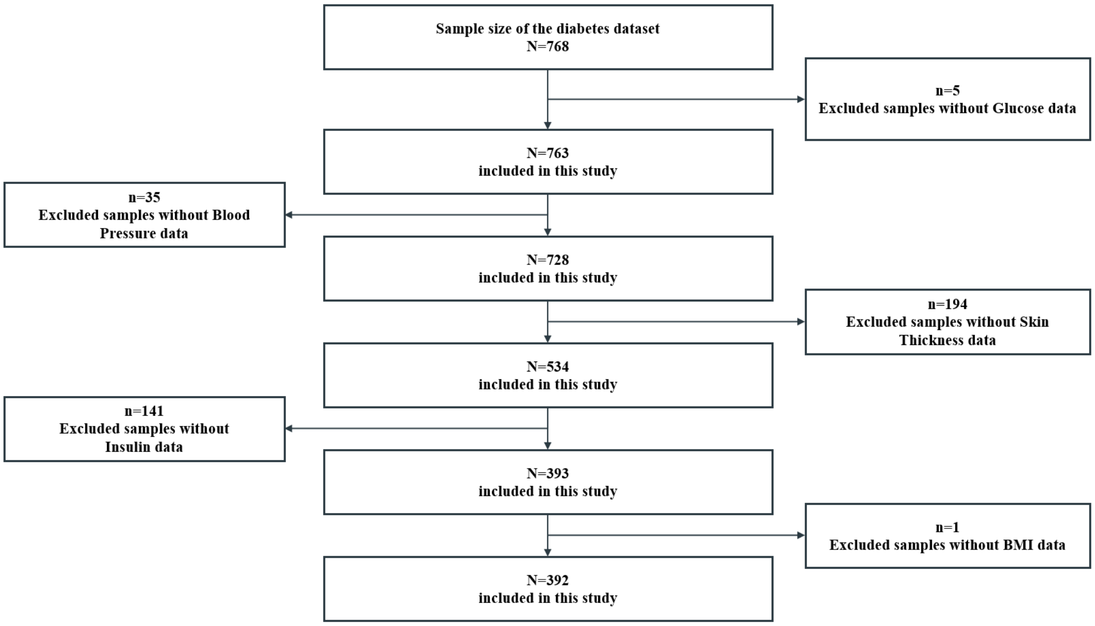
Sample size of the diabetes dataset
N=768
n=5
Excluded samples without Glucose data
N=763
included in this study
n=35
Excluded samples without Blood Pressure data
N=728
included in this study
n=194
Excluded samples without Skin Thickness data
N=534
included in this study
n=141
Excluded samples without Insulin data
N=393
included in this study
n=1
Excluded samples without BMI data
N=392
included in this study

Figure 1: Flow chart of the data cleaning process eligible participants under the selected criteria.

dependent variable. It belongs to the family of generalized linear regression models and is specifically designed for dependent variables that follow a binomial distribution. The binomial distribution arises from a series of repeated trials in a binomial experiment, where each trial has a certain probability of success [14]. The distribution represents the number of successes observed in these trials. Understanding the binomial distribution is crucial for comprehending random phenomena, making effective decisions, and grasping mathematical and probabilistic concepts such as the normal distribution [15].

In the context of binomial regression, the objective is to model the probability of success (typically denoted as "1" in the dataset) as a function of the independent variables. This modeling approach is widely used for predicting binary outcomes due to its simplicity, interpretability, and efficiency. Binomial regression estimates the odds of success, which can be interpreted as the probability of the binary response (e.g., the likelihood of having diabetes). Additionally, it captures the correlation between the predictor variables and the logarithm of the odds of the binary response. By examining the relationships between the independent variables and the log odds, binomial regression allows for understanding the impact of different factors on the likelihood of success in the binary outcome.

2.4.2 Regression Trees. A regression tree is a machine-learning strategy for dealing with regression problems, and it is also known as a decision tree regression. It is a predictive modeling tool that uses a binary tree structure to make predictions based on input features. In a regression tree, the input data is split into subsets based

Table 2: Clinical characteristics of the study samples based on the outcome of diabetes.

Characteristics	Diabetes	Non-diabetes	P Value
Number of subjects	130 (33.2%)	262 (66.8%)	
Age (year)	35.94 (10.63)	28.35 (8.99)	<0.001
Body Mass Index (%)			<0.001
Normal	2 (1.5)	43 (16.4)	
Obesity	110 (84.6)	152 (58.0)	
Overweight	18 (13.8)	67 (25.6)	
Pregnancies	3.00 (6.0)	2.00 (3.0)	<0.001
Glucose	144.50 (47.5)	107.50 (32.0)	<0.001
Blood Pressure	74.00 (15.5)	70.00 (16.0)	<0.001
Skin Thickness	33.00 (13.75)	27.00 (15.75)	<0.001
Insulin	169.50 (111.75)	105.00 (97.75)	<0.001
Diabetes Pedigree Function	0.55 (0.46)	0.41 (0.36)	<0.001

on different feature values. A decision is taken at each node of the tree to partition the data based on a given feature and a related threshold value. This method is continued for each subset recursively until a stopping requirement, such as reaching a maximum depth or a minimum number of samples in a leaf node, is fulfilled. A regression tree's purpose is to reduce the variance of the target variable within each leaf node. The prediction for a new input sample is made by traversing the tree from the root node to a leaf node and taking the average value of the target variable in that leaf node as the predicted output. Regression trees are advantageous because they can handle both numerical and categorical features and can capture nonlinear correlations between input and target variables. They are also interpretable, as the tree structure can be visualized and easily understood. Overall, regression trees are a versatile tool for regression analysis, providing a flexible and interpretable approach to predicting continuous numerical outcomes based on input features.

3 RESULTS AND DISCUSSION

A total of 392 Pima Indian female participants were included after the data cleaning and categorized into 2 groups: diagnosed diabetes (n=130) and non-diabetes (n=262). The mean age was 30.86 years, and the mean of pregnancy frequency was 3.301. Characteristics of the study population among each outcome group were described in Table 2.

From the table, there were significant differences observed between various factors including age, BMI, pregnancies, glucose, blood pressure, skin thickness, insulin, and diabetes pedigree function. The samples without diabetes exhibited certain characteristics compared to those with diabetes. Specifically, individuals without diabetes were younger, with an average age of 28 years compared to 36 years for those with diabetes (P<0.001). Additionally, they had lower pregnancy times, glucose levels, blood pressure, skin thickness, insulin levels, and diabetes pedigree function levels (P<0.001). The side-by-side box plot was made to describe the comparation, which was shown in Figure 2.

The data consists of the number of subjects with percentage or medians with inter quartile ranges. The percentage among participants in different groups were compared using the Chi-square test while the median values among participants in different groups were compared using the Kruskal-Wallis test.

Table 3 displays the results of the univariable and multivariate logistic regression analyses exploring the association between each independent variable and diabetes outcome. The univariable analysis revealed significant associations between all variables included and diabetes outcome (P<0.001). The multivariate analysis showed that increasing age (P=0.065), glucose (P<0.001), BMI (P=0.008), and diabetes pedigree function (P=0.01) were associated with diabetes diagnosis, whereas declining insulin (P=0.01) was also associated with diabetes diagnosis.

By choosing age, BMI, glucose and diabetes pedigree function on the above variables and fitting analysis, a logistic regression model was obtained for predicting Pima Indian female diabetes patients, with P value of all the variables included in this equation no more than 0.01. The AIC value equals to 291.2, which is consistent with the hypothesis above and demonstrates the efficiency of this model. To test its accuracy, a simple cross validation was used by randomly choosing 80% of the sample data as training pool and the 20% remains as testing pool. The accuracy of this model by cross validation is 82.28%.

$$\begin{aligned} \frac{\log p}{1-p} = & -9.771 + 0.0376x_{i,Glucose} + 0.07x_{i,BMI} \\ & +1.333x_{i,DiabetesPedigreeFunction} + 0.037x_{i,Age} \end{aligned} \tag{1}$$

To validate its accuracy further, the logistic regression model was utilized on the expanded dataset. To enhance the effective sample size, the data selection criteria was modified. Samples containing complete data for all four variables were retained, even if that data for other variables were missing. Subsequently, the data cleaning procedure was performed again as shown in Figure 3. As a result, the dataset now includes 752 samples. And from the confusion matrix by fitted model applying on the new dataset, the accuracy of the logistic regression model is 77.48%, with the sensitivity of 92.63%, the specificity of 51.79%, and the kappa value is 0.4793.

To predict the diabetes outcome from another statistics method, regression tree was applied in this dataset. First, R package Caret was applied on the full saturated variables and built a preliminary

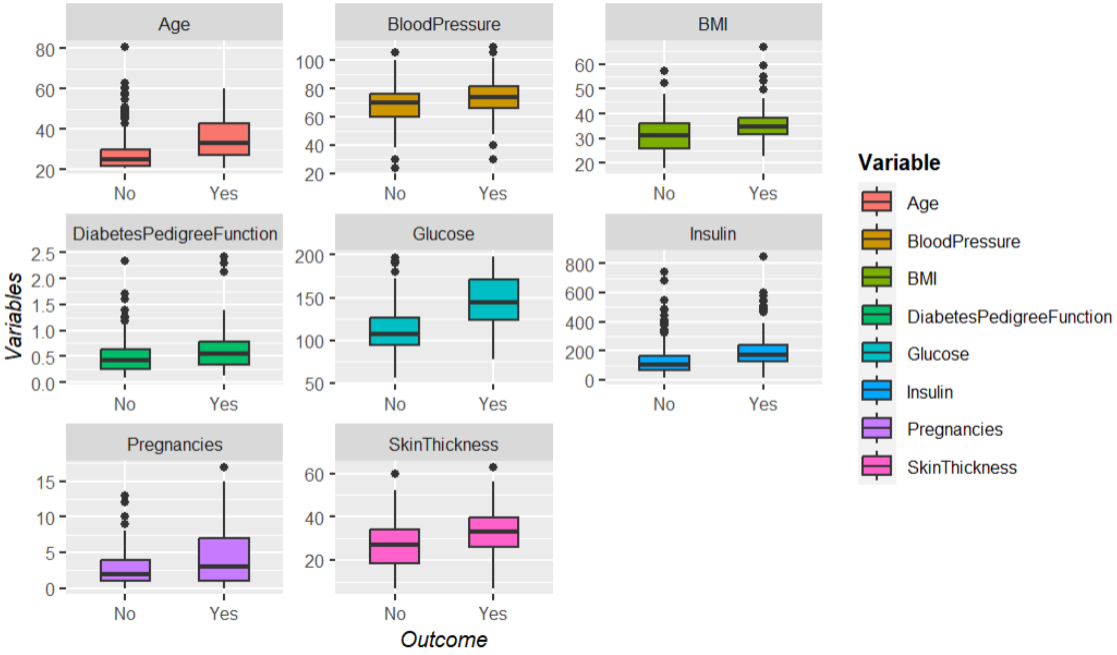

Figure 2: Side-by-side boxplots between diabetes outcome and investigated variables.

Table 3: Univariable and multivariate logistic regression analysis predicting diabetes outcome.

Variable	Univariable Logistic Regression Analysis			Multivariable Logistic Regression Analysis		
	B	SE	P Value	B	SE	P Value
Age (Years)	0.075	0.011	<0.001	3.395e-02	1.838e-02	0.065
Pregnancies	0.167	0.034	<0.001	8.216e-02	5.543e-02	0.138
Glucose	0.042	0.005	<0.001	3.827e-02	5.768e-03	<0.001
Blood Pressure	0.034	0.009	<0.001	-1.420e-03	1.708e-02	0.904
Skin Thickness	0.054	0.011	<0.001	1.122e-02	1.718e-02	0.511
Insulin	0.006	0.001	<0.001	-8.253e-04	1.306e-03	0.010
BMI	0.086	0.017	<0.001	7.054e-02	2.734e-02	0.008
Diabetes Pedigree Function	1.281	0.329	<0.001	1.141e+00	4.274e-01	0.001

B represented as unstandardized regression coefficient; SE represented as standard error of the coefficient; BMI referred as body mass index.

regression tree as shown on Figure 4. It is found that variables actually used in tree construction are age, blood pressure, BMI, diabetes pedigree function, glucose and insulin. Then using complexity Parameter table which contains information about the complexity parameters associated with each node in the tree, the regression tree was pruned aiming to find an optimal trade-off between model complexity and predictive accuracy.

By turning over max depth to 25, a pruned regression tree was optimized as shown on Figure 5 and Figure 6, and Figure 4 is the correlation between max depth and Root Mean Square Error (RMSE). It shows from the pruned regression tree that glucose, BMI, age, diabetes pedigree function and blood pressure were applied in the optimized regression tree model. Then to test its accuracy, cross validation was applied to the regression tree model by randomly choosing 80% of the sample data as training pool and the 20% remains as testing pool. The optimized expanded dataset was further developed by excluding the samples without blood pressure data, which eventually 724 samples. By establishing the confusion matrix, the accuracy of the regression tree model is 76.42%, with the sensitivity of 86.67%, the specificity of 60.42%, and the kappa value is 0.4873, which is consistent to logistic regression above.

In this study, a total of 768 Pima Indian female samples were included. Two prediction methods, logistic regression, and regression trees were used to predict the outcome, and Table 4 showed the accuracy and kappa of the two prediction models, it showed that both prediction models have the similar accuracy which logistic regression showed better accuracy performance, with an accuracy of 77.48% through cross-validation. Key health indicators like age,

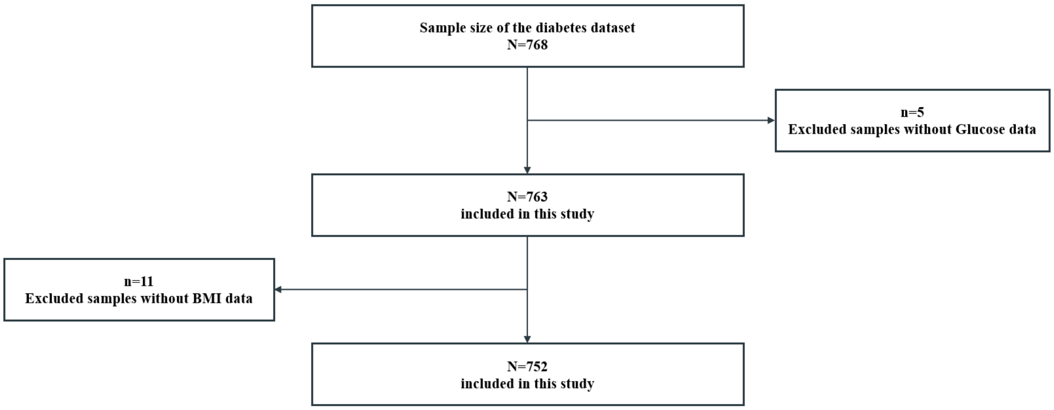
Sample size of the diabetes dataset
N=768
n=5
Excluded samples without Glucose data
N=763
included in this study
n=11
Excluded samples without BMI data
N=752
included in this study

Figure 3: Flow chart of the data cleaning process for expanded dataset.

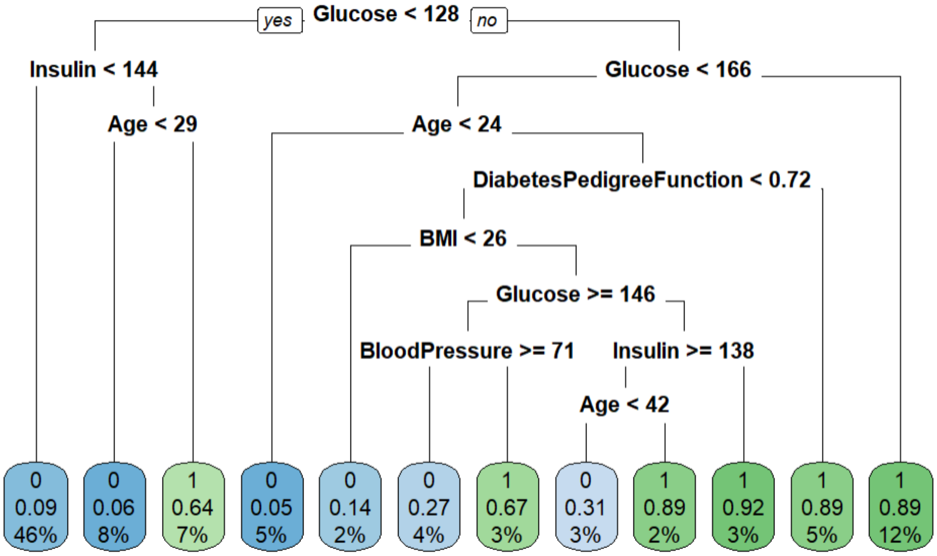

Figure 4: Primary regression tree with full saturate variables.

Table 4: Accuracy and Kappa value of logistic regression and regression tree predicting diabetes outcome

	Logistic Regression	Regression Tree
Accuracy	77.48%	76.42%
Sensitivity	92.63%	86.67%
Specificity	51.79%	60.42%
Kappa	0.4793	0.4873

Accuracy: measures the overall correctness of the predictions and how well the model predicts the correct diabetes outcome; Sensitivity: measures the proportion of true positive predictions out of all actual positive cases in the diabetes sample dataset; Specificity: measures the proportion of true negative predictions out of all actual negative cases in the diabetes sample dataset; Kappa: measures the agreement between predicted and actual diabetes outcome.

body mass index (BMI), glucose level, and diabetes pedigree function were considered as predictors. And from the regression tree, glucose level and BMI might be more strongly associated with the likelihood of having diabetes outcome.

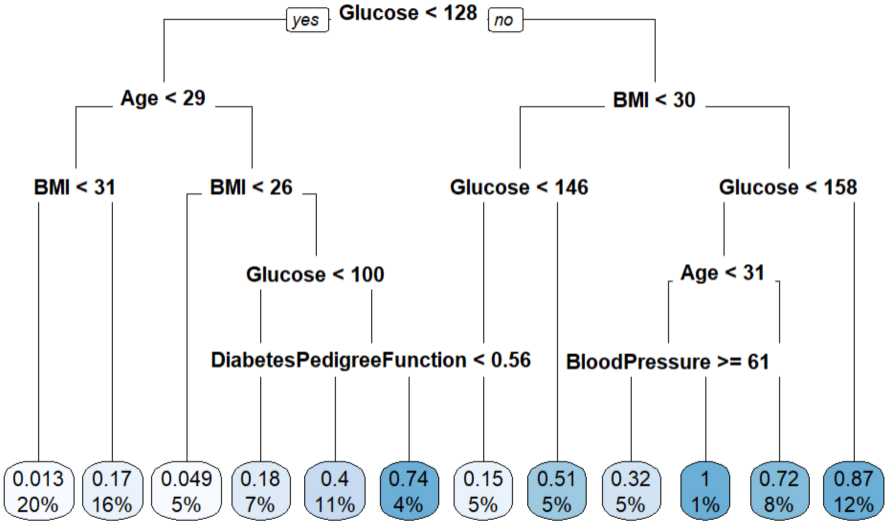

Figure 5: Pruned regression tree model by turning over max depth.

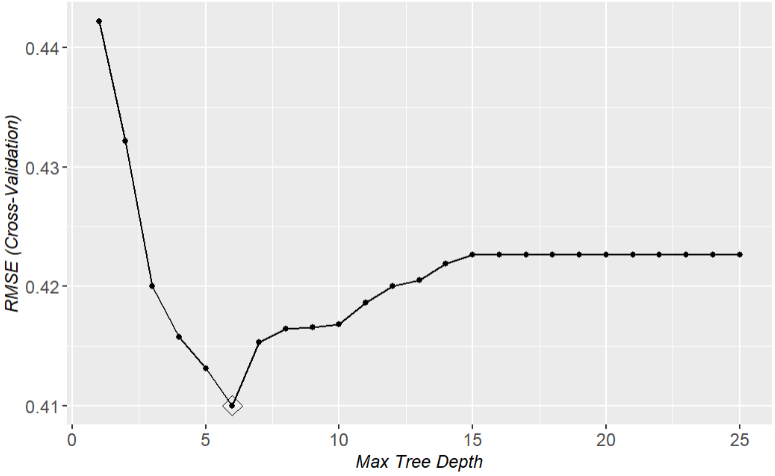

Figure 6: Plot of max tree depth and RMSE.

In order to explain the logistic model, a comprehensive discussion was provided on the main four factors influencing the development of diabetes. The body's regulation of blood glucose levels is impaired in diabetes due to insufficient insulin production or impaired insulin function, resulting in elevated glucose levels in the bloodstream. Higher body mass index (BMI), especially excess abdominal fat, is strongly associated with a higher risk of type 2 diabetes. Disrupted lipid metabolism, ongoing inflammation, and insulin resistance are the root causes of this association. Additionally, due to genetic variations affecting insulin pathways and glucose metabolism, people with a family history of diabetes are at a higher risk. Another important risk factor for diabetes is aging, as metabolic changes brought on by aging, such as decreased insulin sensitivity, decreased insulin production, and altered glucose metabolism, all contribute to the onset of the illness.

The process of diabetes development involves intricate mechanisms starting from the intake of carbohydrates, which are broken down into glucose during digestion. Following absorption into the bloodstream, glucose levels rise, prompting the release of insulin from the pancreas. In individuals with type 2 diabetes, however, cells progressively develop resistance to the action of insulin, impeding glucose uptake. Simultaneously, pancreatic beta cells, responsible for insulin production, may exhibit dysfunction, leading to reduced insulin secretion. Consequently, sustained hyperglycemia ensues, characterized by elevated blood glucose levels. Once hyperglycemia persists and meets established diagnostic criteria, a clinical diagnosis of diabetes is made. This multifaceted condition can give rise to a range of complications and symptoms if not effectively managed. Therefore, optimal diabetes management

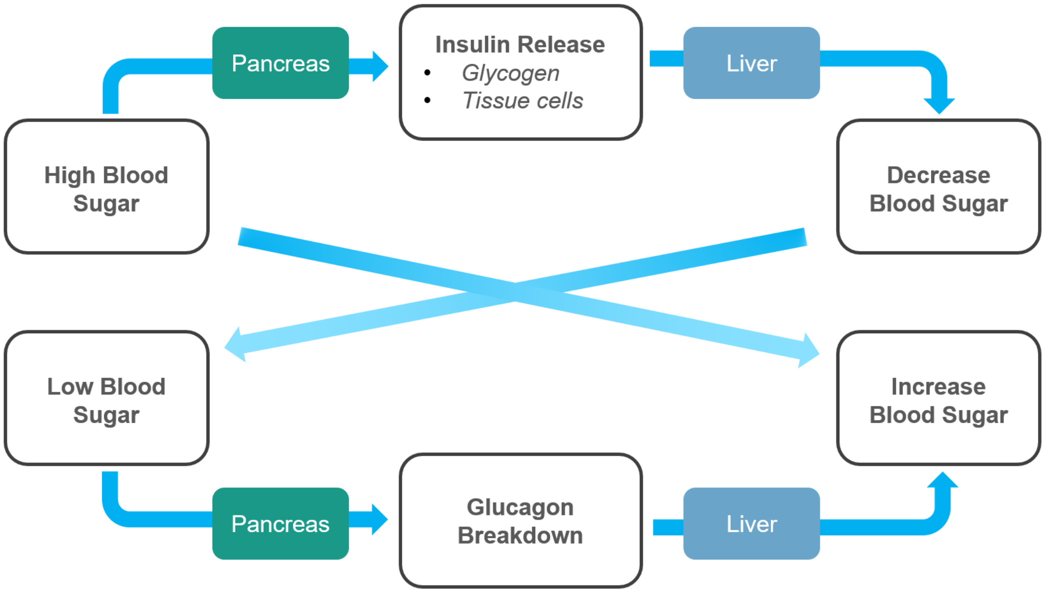

Figure 7: The association between diabetes outcome and glucose.

necessitates lifestyle modifications, pharmacotherapy, and regular monitoring to maintain glycemic control as shown in Figure 7.

The development of diabetes involves complex processes, and one contributing factor is the association between diabetes and BMI (Body Mass Index). BMI is calculated based on an individual's weight and height and serves as an indicator of overall body fatness. Elevated BMI values are associated with a higher risk of developing type 2 diabetes. Excess adipose tissue, especially in the abdominal region, has been linked to insulin resistance, which is a key of type 2 diabetes []. Adipose tissue releases various hormones and inflammatory molecules that can disrupt insulin signaling and hinder glucose uptake by cells. Over time, these metabolic disturbances can result in chronic hyperglycemia, which is the defining feature of diabetes. Lifestyle changes such as adopting a balanced diet, engaging in regular exercise, and achieving and maintaining a healthy weight can help reduce BMI and effectively manage diabetes.

Pregnancies, blood glucose levels, blood pressure, insulin, BMI, function of the diabetes pedigree, and age have all been identified as significant risk factors in relation to diabetes. The likelihood of developing type 2 diabetes is increased by having multiple pregnancies and a history of gestational diabetes. Glucose level, or blood sugar level, plays a critical role as consistently elevated levels over time can lead to complications and organ damage. Due to its role in metabolic and cardiovascular diseases, high blood pressure raises the risk of diabetes. Diabetes is primarily caused by an absolute lack of the hormone insulin, which is involved in the metabolism of glucose. Type 2 diabetes is characterized by insulin resistance or insufficient insulin production. BMI, a measure of body fat, is strongly linked to diabetes, as obesity contributes to insulin resistance and impaired glucose metabolism. The diabetes pedigree function evaluates family history to assess genetic predisposition, as having close relatives with diabetes increases an individual's risk. Age is also a significant factor, with the prevalence of diabetes increasing with advancing age due to age-related physiological changes and lifestyle factors. Understanding and considering these strong risk factors are crucial for identifying individuals at risk for diabetes, implementing preventive measures, and managing the disease effectively.

In addition to the aforementioned risk factors described in the article, there exist other variables that can be quantitatively measured and utilized in the prediction model for diabetes. These variables include lifestyle risks and biochemical parameters, which have shown favorable accuracy in predicting diabetes. A frequently employed approach in the realm of prediction entails the implementation of machine learning models, specifically the decision tree algorithm, which finds extensive usage in scientific investigations. For instance, a study conducted by Meng et al. in 2013 focused on predicting diabetes in a population of Chinese adults in Guangzhou. Logistic regression and decision trees were utilized in this study. By considering factors such as BMI, sleeping duration, physical activities, smoking habits, and alcohol history, the logistic regression model achieved a classification accuracy of 76.13%, with a sensitivity of 79.59% and a specificity of 72.74%. On the other hand, the decision tree model achieved a classification accuracy of 77.87%, with a sensitivity of 80.68% and a specificity of 75.13% [16]. Another study conducted by Esmaily et al. in 2018 incorporated not only demographic characteristics such as age, gender, and BMI but also lifestyle data and biochemical markers including fasted serum triglycerides, total cholesterol, HDL-cholesterol, and LDL-cholesterol. Through the application of decision trees and random forest, this study reported relatively positive results, with the decision tree achieving an accuracy of 64.9%, sensitivity of 64.5%, and specificity of 66.8%, while the random forest achieved an accuracy of 71.1%, sensitivity of 71.3%, and specificity of 69.9% [17]. These studies underscore the significance of other potential variables that are strongly associated with early prediction of diabetes.

Table 5: Comparison analysis of diabetes prediction using PIMA Indian diabetes dataset sources (non-exhaustive)

	Applied Methods	Achieved Accuracy
Gupta et.al (2013)	Decision Tree	81.33%
A.Iyer et.al (2015)	Decision Tree	74.8%
Zou et.al (2018)	Random Forest	77.21%
Joshi et.al (2020)	Logistic Regression, Decision Tree	74%-78%
Shafi et.al (2021)	Naïve Bayes, Decision Tree	73%-76%
Chang et.al (2023)	Naïve Bayes, Random Forest, Decision Tree	74%-80%

For Pima Indian Female diabetes prediction, table 5 showed previous investigations that predict diabetes outcome of Pima Indian Group have applied multiple statistics methods, such as J48 decision tree, Naïve Bayes, and random forest. For example, Joshi et al. achieved accuracy rates ranging from 74% to 78% by employing logistic regression and decision trees, focusing on five key indicators such as the number of pregnancies, glucose levels, pedigree, BMI, and age [18]. Zou et al. successfully predicted diabetes in Pima Indian datasets with an accuracy rate of 77% using the random forest approach [19]. Similarly, Chang et al. achieved a diabetes prediction accuracy range of 74% to 80% by utilizing the Naïve Bayes classifier, random forest classifier, and J48 decision tree models [20]. The accuracy was consistent compared with former investigation. Beyond that, less indicators which can be easily tested were selected to establish a simple diabetes prediction formula based on logistic regression, and two core indicators were shown from decision tree structure. This further investigation was to provide a relatively easy and low-cost method for non-diagnosis Pima Indian Group to establish self-prediction. With the formula outcome that showed a relatively positive diabetes possibilities, early diagnosis and interventions might need to be developed. Furthermore, two core indicators, BMI and blood glucose level, are also critical to diabetes outcome for Pima Indian Group, which might potentially contribute to prevent diabetes for early control management.

The advantage of the study lies in the use of a consistent racial sample comprising female Pima Indian participants. This approach enhances the statistical strength of the study, resulting in more reliable results specifically applicable to Pima Indian females. Additionally, the project had some restrictions even after efforts were made to exclude any potential confounding variables from the investigation of the variables associated with diabetes predictions. As a cross-sectional study, it was particularly challenging to establish a causal relationship between diabetes outcome and variables. Additionally, due to the sparse data and the fact that some variables' values, like blood pressure and glucose, are dynamic rather than static, bias may exist. Besides, more types of variables such as lifestyle data and biochemical markers data would further contribute to the accuracy of prediction model.

4 CONCLUSION

In conclusion, logistic regression showed better accuracy performance, with an accuracy of 77.48% through cross-validation. Key health indicators like age, body mass index (BMI), glucose level, and diabetes pedigree function were considered as predictors. And from the regression tree, glucose level and BMI might be more strongly associated with the likelihood of having diabetes outcome. This research might potentially contribute to simple and low-cost self-prediction of diabetes among Pima Indian females, enabling establishing proactive early prevention strategies and interventions by easily accessible indicators to improve health outcomes for at-risk individuals.

The critical healthcare disparities were faced by the Pima Indians due to limited resources, low income, and unfavorable economic conditions. These challenges have contributed to the delayed detection and management of chronic illnesses, particularly diabetes, leading to adverse health outcomes within the community. However, our study demonstrates the potential of machine learning algorithms as a viable solution to early self . Nevertheless, to ensure the effectiveness and cultural appropriateness of such solutions, collaborative efforts between healthcare providers, researchers, and the Pima community are essential.

REFERENCES

[1] Shao H, Li P, GuroJ, Fonseca V, Shi L, Zhang P. 2022. Socioeconomic factors play a more important role than clinical needs in the use of SGLT2 inhibitors and GLP-1 receptor agonists in people with type 2 diabetes external icon. Diabetes Care, 45(2), 32-33.

[2] Centers for Disease Control and Prevention. 2022. Diabetes Basics. https://www.cdc.gov/diabetes/basics/diabetes.html.

[3] Wonnacott A, *et al.* 2022. MicroRNAs and their delivery in diabetic fibrosis. Advanced drug delivery reviews, 182.

[4] Galaviz K I, *et al.* 2018. Global Diabetes Prevention Interventions: A Systematic Review and Network Meta-analysis of the Real-World Impact on Incidence, Weight, and Glucose. Diabetes care, 41(7), 1526–1534.

[5] Mack R, Tomich P G. 2017. Gestational Diabetes: Diagnosis, Classification, and Clinical Care. Obstetrics and gynecology clinics of North America, 44(2), 207–217.

[6] Nelson R G, *et al.* 1993. Diabetic kidney disease in Pima Indians. Diabetes care, 16(1), 335–341.

[7] Schulz L O, Chaudhari L S. 2015. High-Risk Populations: The Pimas of Arizona and Mexico. Current obesity reports, 4(1), 92–98.

[8] Lillioja S. 1996. Impaired glucose tolerance in Pima Indians. Diabetic medicine: a journal of the British Diabetic Association, 13(9), 127–132.

[9] Szmuilowicz E D, Josefson J L, Metzger B E. 2019. Gestational Diabetes Mellitus. Endocrinology and metabolism clinics of North America, 48(3), 479–493.

[10] Kulshrestha V, Agarwal N. 2016. Maternal complications in pregnancy with diabetes. JPMA. The Journal of the Pakistan Medical Association, 66, 74–77.

[11] Schaefer-Graf U, *et al.* 2018. Diabetes in pregnancy: a new decade of challenges ahead. Diabetologia, 61(5), 1012–1021.

[12] Crandall J P, *et al.* 2008. The prevention of type 2 diabetes. Nature clinical practice. Endocrinology & metabolism, 4(7), 382–393.

[13] Tuomilehto J, Wolf E. 1987. Primary prevention of diabetes mellitus. Diabetes care, 10(2), 238–248.

[14] Prasetyo R B, Kuswanto H, Iriawan N, Ulama B S S. 2020. Binomial regression models with a flexible generalized logit link function. Symmetry, 12(2), 221.

[15] García-García J I, *et al.* 2022. The Binomial Distribution: Historical Origin and Evolution of Its Problem Situations. Mathematics, 10(15), 2680.

[16] Meng XH, *et al.* 2013. Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. Kaohsiung J Med Sci, 29(2), 93–99.

[17] Esmaily, H, *et al.* 2018. A Comparison between Decision Tree and Random Forest in Determining the Risk Factors Associated with Type 2 Diabetes. Journal of

Research in Health Sciences, 18(2), 412.

[18] Joshi, R. D., *et al.* 2021. Predicting Type 2 Diabetes Using Logistic Regression and Machine Learning Approaches. International journal of environmental research and public health, 18(14), 7346.

[19] Zou, Q., *et al.* 2018. Predicting Diabetes Mellitus With Machine Learning Techniques. Frontiers in genetics, 9, 515.

[20] Chang V, Bailey J, Xu Q A, *et al.* 2023. Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms[J]. Neural Computing and Applications, 35(22): 16157-16173.