Scientific Research Publishing

Open Journal of Nursing, 2025, 15(8), 638-665
https://www.scirp.org/journal/ojn
ISSN Online: 2162-5344
ISSN Print: 2162-5336

Statistical Analysis of a Diabetes Dataset and the Impact of Principal Component Analysis on Prediction Accuracy

Elizabeth Diamond*, Faith Idoko, Michael Olowe

Department of Industrial and Systems Engineering, North Carolina A & T State University, Greensboro, NC, USA
Email: *emdiamond@aggies.ncat.edu, foidoko@aggies.ncat.edu, msolowe@aggies.ncat.edu

How to cite this paper: Diamond, E., Idoko, F. and Olowe, M. (2025) Statistical Analysis of a Diabetes Dataset and the Impact of Principal Component Analysis on Prediction Accuracy. *Open Journal of Nursing*, **15**, 638-665.
https://doi.org/10.4236/ojn.2025.158047

Received: May 25, 2025
Accepted: August 19, 2025
Published: August 22, 2025

Abstract

This paper aims to investigate the effectiveness of logistic regression and discriminant analysis in predicting diabetes in patients using a diabetes dataset. Additionally, the paper explores the impact of principal component analysis (PCA) on the prediction accuracy of these methods. The dataset used for this study contains clinical and demographic information of patients with and without diabetes. Logistic regression (LR) and discriminant analysis (DA) were employed to build predictive models using the dataset. The models were then evaluated using various performance metrics such as sensitivity, specificity, and accuracy. The hypothesis (D_0 = patient does NOT have diabetes, whereas D_1 = patient HAS diabetes) is determined with statistical analysis. Results show both logistic regression and discriminant analysis can accurately predict diabetes in patients. Performing PCA did not improve the prediction accuracy of these statistical techniques on the diabetes dataset. The analysis dataset contained 390 patient records with 14 clinical variables. While the dataset provides valuable insights, the relatively small sample size may limit the generalization of the results to broader populations. Our findings suggest that logistic regression or discriminant analysis can be a powerful tool for predicting diabetes in patients, aiding in early detection and effective prevention or management of the disease.

Keywords

Logistic Regression, Discriminant Analysis, Principal Component Analysis

1. Introduction

1.1. Background

According to the Centers for Disease Control and Prevention (CDC), diabetes is

a long-term health condition that impacts how the body converts food into energy. When food is digested, it is transformed into glucose and absorbed into the bloodstream. Insulin is produced by the body to enable the uptake of blood sugar into the cells, where it is used as energy. However, for individuals with diabetes, the body either fails to produce sufficient insulin or is unable to utilize the insulin effectively to meet the energy needs of the body [1].

Recent statistics indicate that 37.3 million people in the United States, which represents 11.3% of the population, have diabetes [2]. Shockingly, one out of every five people with diabetes are unaware of their condition [1]. Diabetes can be classified into three major types, including Type 1, Type 2, and gestational diabetes, which occurs during pregnancy. Any of these variants can affect an individual.

1.2. Problem Description

Diabetes is a global health concern, characterized by the body's inability to regulate blood glucose levels effectively. According to the Centers of Disease Control and Prevention National Diabetes Statistics Report [3], an estimated 130 million adults in the United States have diabetes or prediabetes. Diabetes is the leading cause of kidney failure, limb amputation, and blindness in adults [4]. Over the past few decades, the prevalence of diagnosed type 2 diabetes in adults has risen dramatically, from 4.5% in 1995 to 10.2% in 2020 [4]. The estimated percentage in 2022 has risen to 14.7% [3], and 1.4 million new cases of diabetes were diagnosed among people aged 18 and older in 2019. Adults with a family income below the federal poverty level had the highest prevalence for both men (13.7%) and women (14.4%), while individuals with less education were more likely to have diagnosed diabetes [3].

Prediabetes is a reversible condition where blood sugar levels are high but not as high as those seen in diabetes. Understanding prediabetes is essential as it can prevent or delay the onset of type 2 diabetes. Diet and exercise have a significant influence on diabetes prevention. The National Diabetes Statistics Report contains full details on prediabetes.

The increasing prevalence of diabetes is a significant public health concern, given its associated complications and status as the eighth leading cause of death in the United States. As a result, a considerable amount of research is underway to understand the causes of diabetes and how to prevent it. Our project aims to predict the probability of an individual having diabetes based on various health and physical factors, including cholesterol and glucose levels, systolic and diastolic blood pressure, age, height, weight, body mass index (BMI), waist, hip, and waist-hip ratio [5].

1.3. Project Scope

The objective of this project is to analyze a diabetes dataset and develop a prediction model that can be used to screen for diabetes in patients based on predominant attributes and patterns. This analysis follows the hypothesis testing structure

defined below.

D_0: Diabetes is absent.

D_1: Diabetes is present.

Based on the attributes, the null hypothesis (D_0) states that the patient has no diabetes while the alternate hypothesis (D_1) states that the patient actually has diabetes.

Three types of multivariate statistical techniques—principal component analysis, discriminant analysis, and logistic regression analysis will be employed. Publicly available data on diabetes will be used for this analysis. The following research questions serve as a guide for this study.

- Do the attributes in the dataset provide significant information on the presence of diabetes in a patient?
- How well do Discriminant analysis and Logistic regression predict the presence of diabetes in a patient?
- Does the use of dimensionality techniques such as principal component analysis provide better attributes and prediction results?

The remaining part of this article is organized as follows: Section 2 includes previous work that addressed the same problem statements already discussed. Section 3 introduces the details of the used dataset, the pre-processing and data preparation phase and the machine learning algorithms used. Further, the results of statistical techniques used, and the associated accuracy are presented and discussed in Section 4. We finish with a conclusion in Section 5.

2. Literature Review

The prediction of diabetes is an important topic in medical research and has been the subject of numerous studies. In Table 1, we have highlighted major findings by researchers who have adopted several statistical procedures in the bid to predict diabetes. Some of these techniques include principal component analysis, logistic regression, decision trees, support vector machines (SVM), discriminant analysis and artificial neural networks (ANN) [6].

Table 1. Literature review details.

S/N	Technical Paper	Year of Publication	Statistical Techniques	Summary of Findings
1	Zhu, Changsheng, Christian Uwa Idemudia, and Wenfang Feng. "Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques." Informatics in Medicine Unlocked 17 (2019): 100179. [7]	2019	Logistic Regression	They determined ways of improving the k-means clustering and logistic regression accuracy result of Pima Indian Diabetes dataset using PCA (principal component analysis), k-means and logistic regression algorithm. The experimental investigations show that PCA improved the k-means clustering algorithm and logistic regression classifier accuracy versus the result of other published studies, with a k-means output of 25 more correctly classified data, and a logistic regression accuracy of 1.98% higher.

Continued

2	Rajendra, Priyanka, and Shahram Latifi. "Prediction of diabetes using logistic regression and ensemble techniques." Computer Methods and Programs in Biomedicine Update 1 (2021): 100032. [8]	2021	Logistic Regression	Logistic Regression technique was used to develop a prediction model using two datasets (Pima & Vanderbilt) and two ensemble methods were further employed to improve the model performance by producing better predictions compared to a single model. With the adoption of ensemble methods, performance of the model was found to increase to 78% for dataset 1 and 93% for dataset 2.
3	Joshi, Ram D., and Chandra K. Dhakal. "Predicting type 2 diabetes using logistic regression and machine learning approaches." International journal of environmental research and public health 18.14 (2021): 7346. [9]	2021	Logistic Regression/ Decision Tree	By employing the use of Logistic Regression and Decision Tree, the researchers sought to predict type 2 diabetes for Pima Indian women. Their analysis found five predominant predictors of type 2 diabetes which are glucose, pregnancy, body mass index (BMI), diabetes pedigree function, and age.
4	Polat, Kemal, Salih Güneş, and Ahmet Arslan. "A cascade learning system for classification of diabetes disease: Generalized discriminant analysis and least square support vector machine." Expert systems with applications 34.1 (2008): 482-487. [10]	2008	Discriminant Analysis/ SVM	They developed a cascade learning system for accurately predicting diabetes in patients using a combined model called GDA-LS-SVM (Generalized Discriminant Analysis and Least Square Support Vector Machine). This model gave a classification accuracy relatively higher than the conventional LS-SVM model.
5	Alharan, Abbas FH, *et al.* "Improving classification performance for diabetes with linear discriminant analysis and genetic algorithm." 2021 Palestinian International Conference on Information and Communication Technology (PICICT). IEEE, 2021. [11]	2021	Discriminant Analysis/ Genetic Algorithm	The researcher employed Discriminant Analysis and Genetic Algorithm on two datasets (Pima Indian Diabetes and Data of Dr John Schorling). They used the two techniques for feature selection and four techniques were used for classification. They found that Random Forest classifiers using DA and GA gave accuracy up to 91%.
6	Abdulhadi, Nour, and Amjed Al-Mousa. "Diabetes detection using machine learning classification methods." 2021 International Conference on Information Technology (ICIT). IEEE, 2021. [6]	2021	Linear Discriminant Analysis/ Random Forest	The researchers were able to predict type 2 diabetes on female patients using the Pima Indian Diabetes Dataset obtained from National Institute of Diabetes and Digestive and Kidney Diseases were used. The Random Forest Classifier gave an accuracy of 82%.
7	El_Jerjawi, Nesreen Samer, and Samy S. Abu-Naser. "Diabetes prediction using artificial neural network." (2018). [12]	2018	Artificial Neural Network	The researchers used artificial neural networks to predict whether a person is diabetic or not. The criterion was to minimize the error function in neural network training using a neural network model. After training the ANN model, the average error function of the neural network was equal to 0.01 and the accuracy of the prediction of whether a person is diabetic or not was 87.3%.
8	Srivastava, Suyash, *et al.* "Prediction of diabetes using artificial neural network approach." Engineering Vibration, Communication and Information Processing: ICoEVCI 2018, India. Springer Singapore, 2019. [13]	2019	Artificial Neural Network	In this research work, a data sample of Pima Indians was taken to predict the possibility of diabetes. Among several algorithms of Machine learning, Artificial Neural Network (ANN) was chosen for building the model to predict diabetes. This model gave a prediction accuracy of 92% with the possibility of achieving higher accuracy if trained with a larger dataset.

Table 2 highlights some of the pros and cons of using these identified statistical techniques based on findings.

Table 2. Pros and cons of statistical techniques.

Statistical Technique	Advantages	Disadvantages
Logistic Regression	- Simple and easy to interpret - Can handle categorical and continuous predictors - Can model the probability of an outcome	- Assumes linear relationship between predictors and outcome - Assumes independence of observations - May not handle complex interactions well
Decision Trees	- Nonlinear relationships between predictors and outcome can be captured. - Can handle both categorical and continuous predictors - Easy to interpret and visualize	- Prone to overfitting, especially with noisy data - Unstable: small changes in data can result in large changes in tree structure - Tendency to favor predictors with many categories
Random Forests	- Nonlinear relationships between predictors and outcome can be captured - Can handle both categorical and continuous predictors - Reduces overfitting by combining multiple decision trees - Can handle missing data	- Less interpretable than decision trees - Requires larger sample size and longer training time than decision trees - Can be computationally intensive
Support Vector Machines	- Can model complex, nonlinear relationships between predictors and outcome - Effective for high-dimensional data - Can handle both categorical and continuous predictors	- Requires careful selection of hyperparameters - Can be sensitive to outliers - Limited interpretability
Neural Networks	- Can model complex, nonlinear relationships between predictors and outcome - Can handle both categorical and continuous predictors - Can learn from noisy or incomplete data	- Requires larger sample size and longer training time than other methods - Can be computationally intensive - Less interpretable than other methods
Discriminant Analysis	- Can handle multiple predictors and multiple outcome classes - Assumes normality and equal covariance matrices among predictors - Can provide insight into which predictors are most important - Can handle missing data	- Assumes linear relationships between predictors and outcome - Sensitive to outliers and non-normality of data - Limited to two outcome classes (linear discriminant analysis) or assumes equal prior probabilities (quadratic discriminant analysis)

2.1. Diabetes Logistic Regression Use Case

A multivariate logistic regression equation can be used to screen for diabetes. In November of 2002, Diabetes Care published an article describing the development and validation of this empirical equation. The research design and methods were unique as subjects were from international sources. A predictive equation was developed with data collected from 1032 Egyptian subjects with no history of diabetes. The equation incorporated age, gender, BMI, postprandial time, and random capillary plasma glucose as independent covariates for prediction of undiagnosed

diabetes. These covariates were based on a fasting plasma glucose level of ≥126 mg/dl and/or a plasma glucose level 2 hr after a 75-g oral glucose load ≥200 mg/dl. The equation was validated using data collected from an independent sample of 1065 American subjects [14]. The predictive equation was calculated with the following logistic regression parameters:

$P = 1/(1 - e^{-x})$, where $x = -10.0382 + [0.0331*(\text{age in years}) + 0.0308*(\text{random plasma glucose in mg/dl}) + 0.2500*(\text{postprandial time assessed as 0 to} \geq 8 \text{ hr}) + 0.5620*(\text{if female}) + 0.0346*(\text{BMI})]$.

The cut-off point for the prediction of previously undiagnosed diabetes was defined as a probability value ≥ 0.20. The equation's sensitivity was 65%, specificity 96%, and positive predictive value (PPV) 67%. When applied to a new sample, the equation's sensitivity was 62%, specificity 96%, and PPV 63% [14]. The equation improved on recommended methods of screening for undiagnosed diabetes and could be easily implemented in a handheld programmable calculator to predict previously undiagnosed diabetes [14]. Machine learning, using several different models, has been attempted without further model improvement [15].

2.2. Performance Metrics

The sensitivity of a test is defined as the proportion of people *with disease* who will have a *positive* result. A test with a high sensitivity is useful for "ruling out" a disease if a person tests negative [17].

The specificity of a test is the proportion of people *without* the disease who will have a *negative* result. A test with a high specificity is useful for "ruling in" a disease if a person tests positive [17].

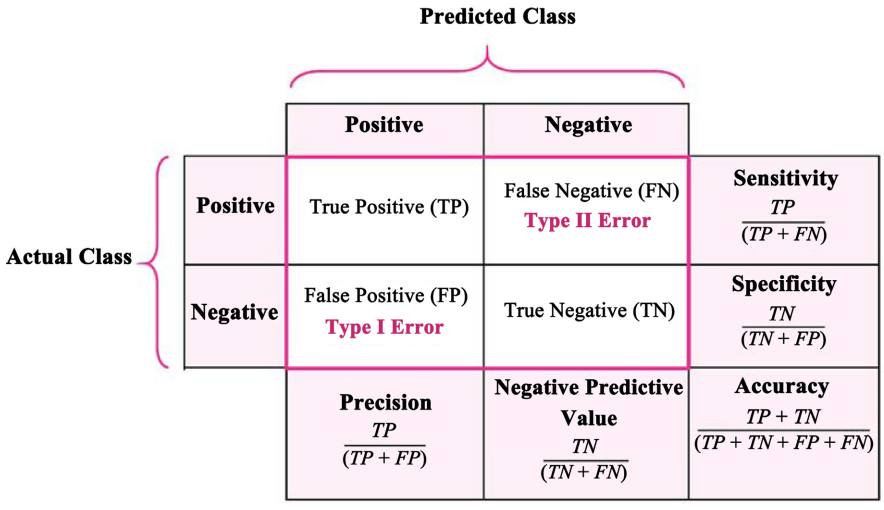

Figure 1. Sensitivity, specificity and more (image source [16]).

The predictive value of a test is determined by the test's sensitivity and specificity and by the prevalence of the condition for which the test is used. Both PPV and NPV vary with changing prevalence of disease. It will therefore be wrong for clinicians to directly apply published predictive values of a test to their own popula-

tions when the prevalence of disease in their population is different from the prevalence of disease in the population in which the published study was carried out [17]. **Figure 1** shows a graphical representation of how Sensitivity, Specificity, Negative predictive value (NPV), Positive predictive value (PPV), and Accuracy are calculated.

3. Methodology

Data Collection and Description

For this work, publicly available data from Kaggle was used, which was initially sourced from the National Institute of Diabetes and Digestive and Kidney Diseases. The dataset contains 390 instances with 14 predictor variables on patients, including their overall cholesterol level, HDL cholesterol level, cholesterol to HDL ratio, glucose levels, systolic and diastolic blood pressure, age, height, weight, BMI, waist and hip measurements, and diabetes status. All variables and their meanings are presented in **Table 3**. These values represent the ground truth labels in the dataset used for model training and evaluation. Performance metrics were chosen to provide a comprehensive evaluation of model performance, including sensitivity (recall), specificity (true negative rate), precision (positive predictive value), negative predictive value, and overall accuracy. These metrics were selected to reflect the clinical importance of both correctly identifying diabetic patients and minimizing false positives. In addition, a scatter plot matrix, and a correlation matrix were plotted to check for multicollinearity and to identify whether any strongly correlated features were present in the dataset that could impact model performance.

The dependent variable in this analysis was the binary variable Outcome, representing the presence or absence of diabetes, while the 13 independent variables were the diagnostic measures mentioned above. Before analysis, the data were cleaned to ensure their suitability for the study.

The National Institute of Diabetes and Digestive and Kidney Diseases collects data from different individuals through various programs and sources, such as surveys, the Diabetes Prevention Program, the National Diabetes Information Clearinghouse, and clinical trials. For example, the National Health and Nutrition Examination Survey (NHANES), conducted by the Centers for Disease Control and Prevention (CDC) in collaboration with the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), collects information on diabetes prevalence, risk factors, and treatment. In contrast, the National Diabetes Prevention Program (NDPP) collects data on participants' demographics, medical history, physical activity, and dietary habits to evaluate the program's effectiveness. The National Diabetes Information Clearinghouse (NDIC) collects and compiles data on diabetes incidence, prevalence, risk factors, and complications from scientific literature and government reports. Finally, NIDDK conducts and supports clinical trials to test new treatments and interventions for diabetes and related conditions, collecting data on participants' demographics, medical history, symp-

toms, and treatment outcomes to evaluate the safety and effectiveness of the intervention.

Table 3. Data description.

S/N	Attribute Name	Attribute Description	Unit
1	Age	The patient's age	years
2	Weight	The patient's weight in kilograms	pounds (lbs)
3	BMI	The patient's body mass index, calculated as weight divided by height squared	lbs/m^2
4	Systolicbp	The patient's systolic blood pressure in millimeters of mercury	mmHg
5	Diastolicbp	The patient's diastolic blood pressure level	mmHg
6	Waist	The patient's waist circumference	Inches
7	Hip	The patient's hip circumference	Inches
8	Height	The patient's height	Inches
9	Waisthip ratio	The patient's waist to hip ratio	
10	Cholesterol	The patient's cholesterol level in milligrams per deciliter	mg/dL
11	Glucose	The patient's glucose level	mg/dL
12	hdlchol	The patient's HDL (High-density lipoprotein) cholesterol level	mg/dL
13	cholhdlratio	The patient's total cholesterol to HDL cholesterol ratio	
14	Diabetes	A binary variable indicating whether the patient has been diagnosed with diabetes (0 = no, 1 = yes).	

In this paper, we aim to predict the susceptibility of patients to diabetes using multivariate statistical techniques. Logistic Regression (LR) and Discriminant Analysis (DA) were chosen due to their established use in the field of medicine and engineering with categorical or discrete response variables. They are also robust to outliers and missing data, making them suitable for real-world datasets that may be incomplete or contain errors.

The first phase of our research involves using DA and LR models to determine the dimensions of the cleaned Kaggle diabetes dataset that can reliably and accurately classify subjects into groups (*i.e.*, diabetic or non-diabetic). LR will be used to identify a linear combination of independent variables (IVs) that best predicts membership in the diabetic or non-diabetic group, as measured by a categorical dependent variable (DV). This will involve deriving an equation for predicting the likelihood of diabetes in identified subjects. Additionally, we will fit the DA model to our dataset to derive the two discriminant functions.

In the second phase of our research, we will perform Principal Component Analysis (PCA) on our dataset and fit the LR and DA models again on the PCA-

reduced dataset. We will then check the model fit and accuracy of the LR and DA models on both the original and PCA-reduced datasets to draw insights into the prediction accuracies [8]. Finally, we will use the K-fold cross-validation technique to compare the performance of the LR and DA models [18]. We used SAS software and Python IDE as the statistical tools for coding. These statistical methods will help us determine the most accurate procedures for predicting diabetes susceptibility and provide insights into the usefulness of PCA as a tool for data reduction in medical studies [7]. Principal Component Analysis (PCA) was applied as an exploratory technique for dimensionality reduction. The variance explained by each component was examined through the proportion of the corresponding eigenvalues. The scree plot was used to determine the optimal number of principal components to retain for the model. Each principal component is a linear combination of the original features and the components are orthogonal (uncorrelated), allowing them to represent distinct directions of variance in the data. These components were used for model training to evaluate whether dimensionality reduction would improve classification performance. The component loadings were examined using the Maximum Likelihood Estimates from logistic regression and the Linear Discriminant Function coefficients.

Model development for both Linear Discriminant Analysis (LDA) and Logistic Regression was performed using SAS software. The LDA model was developed based on the statistical analysis of linear discriminant functions, which aim to find a linear combination of features that best separates the two classes (diabetic and non-diabetic). The Logistic Regression model was developed as a linear model that estimates the probability of diabetes occurrence based on the input features through a logistic function. Both models were trained on the original set of 14 features and, separately, on a reduced set of principal components. Training was performed on both the original feature set and the reduced PCA-based feature set to enable direct comparison of model performance between the full and PCA-based representations.

4. Results and Discussion

4.1. Model Adequacy Check

To ensure the validity of our analysis, several data pre-processing steps were conducted before assessing the adequacy of our model. Missing data, outliers, and unreadable characters in our dataset were checked for and removed. Also, special characters, such as commas and semicolons were fixed, to ensure the data was properly formatted.

Next, several tests (in SAS) were conducted to ensure that our dataset met the necessary conditions for conducting multivariate statistical analysis, specifically Logistics Regression (LR) and Discriminant Analysis (DA). As multivariate analysis relies on assumptions of normality and homogeneity of variance, we performed univariate normality checks on randomly selected variables in the dataset to verify that the data followed a normal distribution.

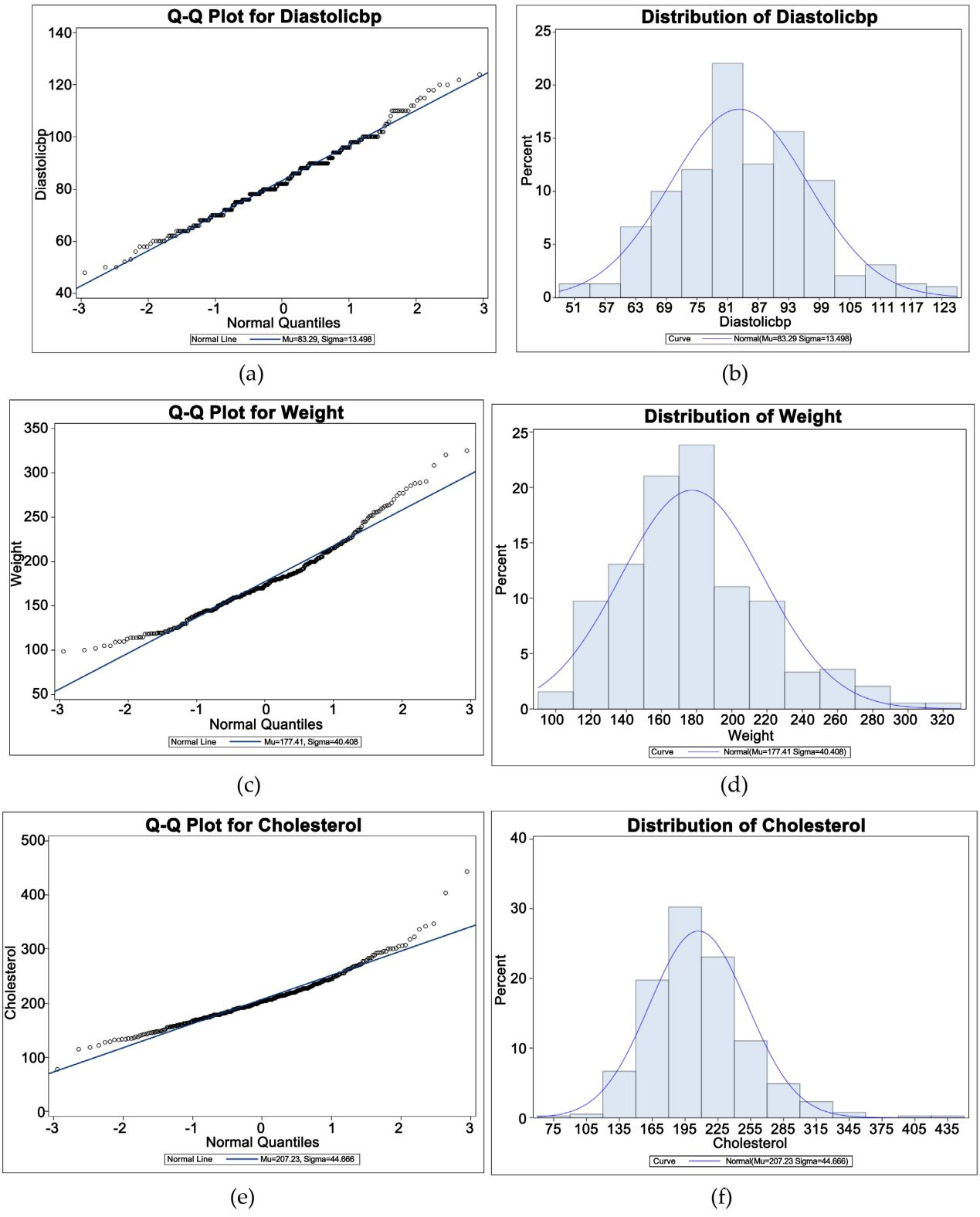

Figure 2. (a-f) Q-Q plots and distribution curves of randomly selected variables.

Due to the complexity involved in testing multivariate normality, a univariate normality procedure was opted for, instead. Investigations showed that the dataset

met the necessary normality assumptions, as confirmed by the Q-Q plots and distribution curves presented in Figure 2. These steps ensured that our dataset was suitable for further statistical analysis.

Normality Test/Test of Homoscedasticity

Next, the test of homoscedasticity (homogeneity of variance) was conducted in SAS. This is given in the scatter matrix provided in Figure 3. No known pattern or trend can be seen, implying that the test of homoscedasticity is not violated.

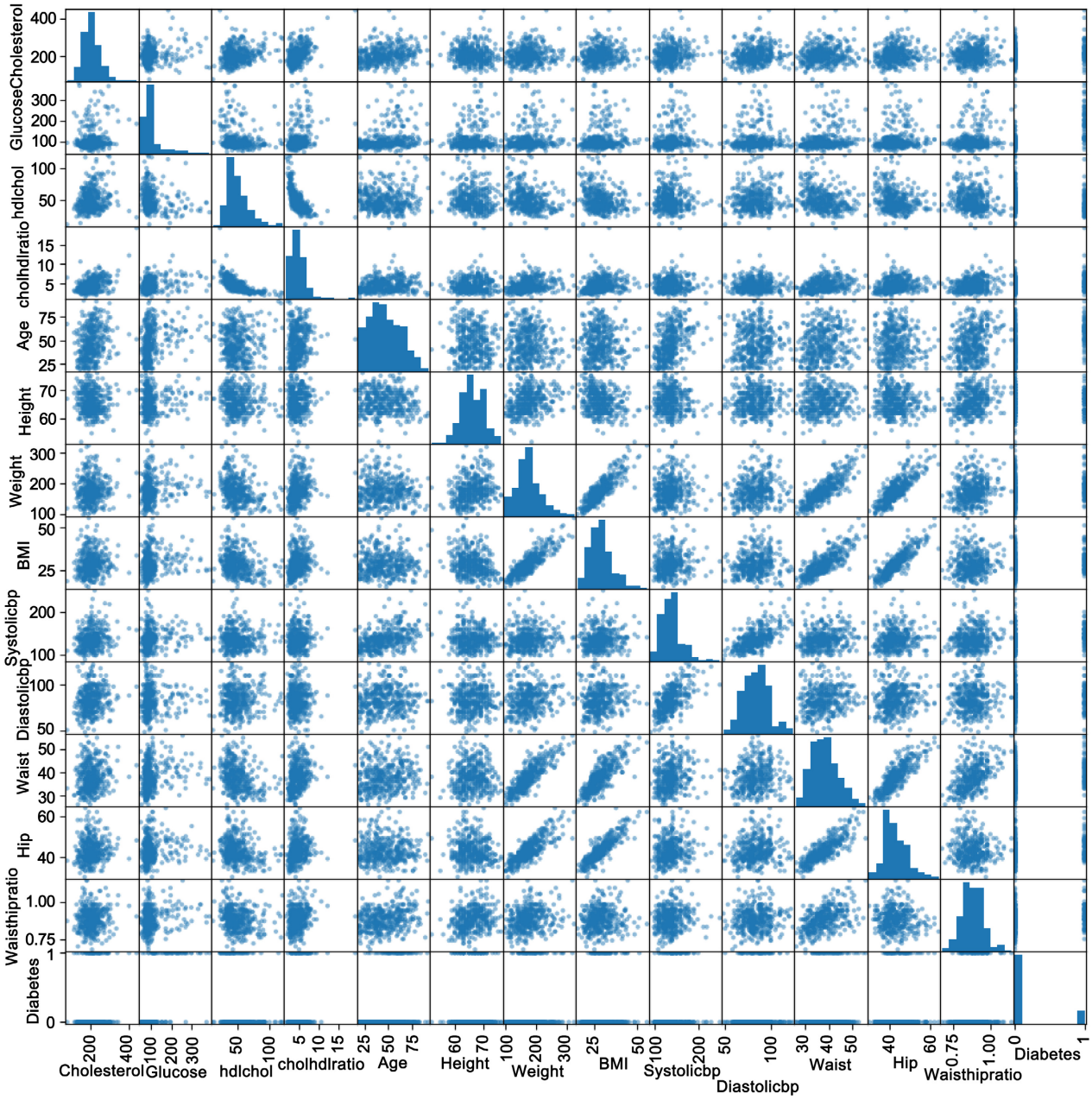

Figure 3. Scatter plot matrix.

The first step involved conducting a correlation analysis amongst each of the

independent variables. From the results obtained and presented in Table 4, there are no such variables with very high correlation among the independent variables. Although, high correlation was recorded between Waist and Weight (**0.8478**), Hip and Weight (**0.8270**) and Waist and Hip (**0.8352**). We will still proceed with our analysis.

Table 4. Correlation matrix.

Correlation Matrix													
	Cholesterol	Glucose	hdlchol	Cholhdl-ratio	Age	Height	Weight	BMI	Systolicbp	Diastolicbp	Waist	Hip	Waisthip-ratio
Cholesterol	1.0000	0.1581	0.1932	0.4759	0.2473	−0.0636	0.0624	0.0917	0.2077	0.1662	0.1340	0.0934	0.0918
Glucose	0.1581	1.0000	−0.1583	0.2822	0.2944	0.0981	0.1904	0.1293	0.1628	0.0203	0.2223	0.1382	0.1851
hdlchol	0.1932	−0.1583	1.0000	−0.6819	0.0282	−0.0872	−0.2919	−0.2419	0.0318	0.0783	−0.2767	−0.2238	−0.1588
Cholhdl-ratio	0.4759	0.2822	−0.6819	1.0000	0.1632	0.0812	0.2788	0.2284	0.1155	0.0382	0.3133	0.2089	0.2433
Age	0.2473	0.2944	0.0282	0.1632	1.0000	−0.0822	−0.0568	−0.0092	0.4534	0.0686	0.1506	0.0047	0.2752
Height	−0.0636	0.0981	−0.0872	0.0812	−0.0822	1.0000	0.2554	−0.2596	−0.0407	0.0436	0.0574	−0.0959	0.2525
Weight	0.0624	0.1904	−0.2919	0.2788	−0.0568	0.2554	1.0000	0.8601	0.0975	0.1665	0.8478	0.8270	0.2505
BMI	0.0917	0.1293	−0.2419	0.2284	−0.0092	−0.2596	0.8601	1.0000	0.1214	0.1453	0.8107	0.8817	0.1009
Systolicbp	0.2077	0.1628	0.0318	0.1155	0.4534	−0.0407	0.0975	0.1214	1.0000	0.6037	0.2109	0.1553	0.1379
Diastolicbp	0.1662	0.0203	0.0783	0.0382	0.0686	0.0436	0.1665	0.1453	0.6037	1.0000	0.1658	0.1439	0.0779
Waist	0.1340	0.2223	−0.2767	0.3133	0.1506	0.0574	0.8478	0.8107	0.2109	0.1658	1.0000	0.8352	0.5142
Hip	0.0934	0.1382	−0.2238	0.2089	0.0047	−0.0959	0.8270	0.8817	0.1553	0.1439	0.8352	1.0000	−0.0377
Waisthip-ratio	0.0918	0.1851	−0.1588	0.2433	0.2752	0.2525	0.2505	0.1009	0.1379	0.0779	0.5142	−0.0377	1.0000

4.2. Descriptive Statistics

As can be seen from the result of the simple statistics conducted on the Diabetes dataset (please see Table 5), the mean age of the patients is 47 years, which implies that the population sampled is an aging population with a standard deviation of approximately 16.4 years. The average height and weight of the population are 66 inches and 177.4 lbs respectively (corresponding standard deviations were measured as 3.9 inches and 40.4 lbs respectively). The systolic/diastolic blood pressure averaged 137/83. According to the American Heart Association, the systolic to diastolic blood pressure is ideal around 120/80, although this could vary depending on age, sex and health condition. The glucose level average for all the patients was 107.3 mg/dL but the dispersion or spread was relatively high (standard deviation = 53.80 mg/dL). The descriptive statistics are also shown in the visualization plot (error plot) shown in Figure 4.

Table 5. Simple statistics of the dataset.

Simple Statistics													
	Cholesterol	Glucose	hdlchol	Cholhdl-ratio	Age	Height	Weight	BMI	Systolic-bp	Diastolic-bp	Waist	Hip	Waisthip-ratio
Mean	207.23	107.34	50.27	4.52	46.77	65.95	177.41	28.78	137.13	83.29	37.87	42.99	0.88
StD	44.67	53.80	17.28	1.74	16.44	3.92	40.41	6.60	22.86	13.50	5.76	5.66	0.07

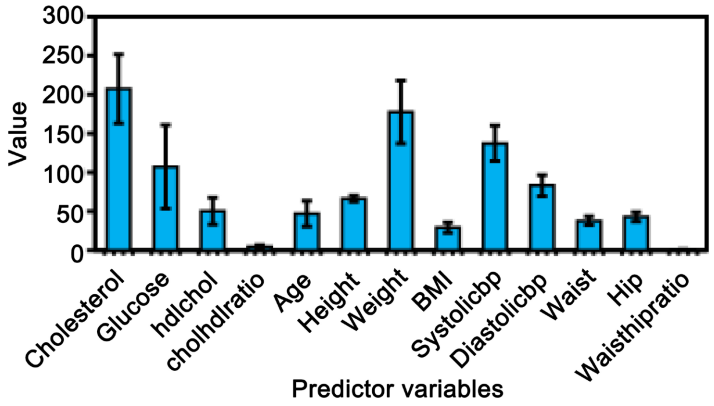

Figure 4. Error plots of the predictor variables.

4.3. Logistic Regression Results

The choice of Logistic Regression (LR) was motivated by the fact that the dependent variables in our dataset were categorical (binary). We fitted the pre-processed dataset to a Logistic Regression model and checked the output file for the final results in SAS.

To assess the validity and goodness of fit of our model, the −2LogL (where L is the likelihood function), AIC (Akaike Information Criterion), and SC (Schwarz's Bayesian Criterion) values were evaluated and shown in Table 6. Lower values of −2LogL, AIC, and SC typically indicate better model fitting, and we will discuss this in more detail in another section.

The Chi-square test statistics yielded p-values of ≤0.001, which is less than the level of significance of 0.05, indicating that the model is statistically significant.

Table 6. Model fit statistics.

Model Fit Statistics			
Criterion	**Intercept Only**	**Intercept and Covariates**	
AIC	336.87	191.24	
SC	340.84	246.76	
−2 Log L	334.87	163.24	
Testing Global Null Hypothesis: BETA = 0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	171.64	13	<0.0001
Score	194.80	13	<0.0001
Wald	66.21	13	<0.0001

Overall, these results suggest that our fitted Logistic Regression model fits our dataset well and provides a statistically significant explanation for the relationship between our independent and dependent variables. A subsequent section will provide further details on the model performance and interpretation. The Logistic Regression model below is formulated according to the SAS output file shown:

$$\ln\left(\frac{p}{1-p}\right) = -11.4431 + 0.0121 * \text{cholesterol} + 0.0372 * \text{glucose} - 0.0278 * \text{hdlchol} - 0.1552 * \text{cholhdlratio} + 0.0315 * \text{Age} + 0.0392 * \text{Height} - 0.0174 * \text{Weight} + 0.1258 * \text{BMI} + 0.00639 * \text{Systolicbp} + 0.00868 * \text{Diastolicbp} + 0.1098 * \text{Waist} - 0.08333 * \text{Hip} - 2.6863 * \text{Waisthipratio}$$

where p is the probability of a patient having diabetes based on the maximum likelihood estimate approach (MLE). From Table 7, the p-value of glucose being less than 0.05 indicates that the glucose level is a predominant predictor variable from the LR model. The medical implication of this result is that higher glucose levels of patients could be a pointer to a higher risk of diabetes.

Table 7. Maximum likelihood.

Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	−11.4431	26.1632	0.1913	0.6618
Cholesterol	1	0.0121	0.00856	1.9913	0.1582
Glucose	1	0.0372	0.00544	46.7283	<0.0001
hdlchol	1	−0.0278	0.0274	1.0318	0.3097
Cholhdl-ratio	1	−0.1552	0.2601	0.3559	0.5508
Age	1	0.0315	0.0171	3.3992	0.0652
Height	1	0.0392	0.2513	0.0243	0.8761
Weight	1	−0.0174	0.0433	0.1616	0.6876
BMI	1	0.1258	0.2538	0.2456	0.6202
Systolicbp	1	0.00639	0.0121	0.2794	0.5971
Diastolicbp	1	0.00868	0.0208	0.1744	0.6762
Waist	1	0.1098	0.5561	0.0390	0.8435
Hip	1	−0.0833	0.4790	0.0302	0.8620
Waisthip-ratio	1	−2.6864	24.7747	0.0118	0.9137

Additionally, the Odds ratio estimates other than 1 indicates that the independent variables have some form of relationship with the outcome variables. The Odds ratio estimate is shown in Table 8.

Table 8. Odds ratio estimates.

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
Cholesterol	1.012	0.995	1.029
Glucose	1.038	1.027	1.049
hdlchol	0.973	0.922	1.026

Continued

Cholhdl-ratio	0.856	0.514	1.426
Age	1.032	0.998	1.067
Height	1.040	0.636	1.702
Weight	0.983	0.903	1.070
BMI	1.134	0.690	1.865
Systolicbp	1.006	0.983	1.031
Diastolicbp	1.009	0.968	1.051
Waist	1.116	0.375	3.319
Hip	0.920	0.360	2.353
Waisthip-ratio	0.068	<0.001	>999.999

Given the results above, the following inferences can be deduced:

1) The odds ratio of 1.134 for BMI suggests that the odds of having diabetes are about 13.4% higher in obese individuals compared to non-obese individuals.

2) The odds ratio of 1.032 for Age suggests that the odds of having diabetes are about 3.2% higher in older individuals compared to younger individuals.

3) The odds ratios of 1.006 and 1.009, respectively, for Systolic and Diastolic blood pressure suggest that the odds of having diabetes are almost the same (about 0.6% and 0.9% higher) in groups with high and low blood pressure.

4) The odds ratio of 1.038 for glucose levels suggests that the odds of having diabetes are about 3.8% higher in individuals with higher glucose levels compared to those with lower glucose levels.

Overall, the odds ratios provide insight into the relationship between each independent variable and the dependent variable (diabetes) in the logistic regression model. These interpretations can be used to better understand the factors that are associated with the presence of diabetes in the population under study.

4.4. Discriminant Analysis Results

Next, we performed a Discriminant Analysis (DA) on the same pre-processed Kaggle diabetes dataset that was used for the Logistic Regression (LR) model. DA is a similar statistical procedure to LR, and it was carried out using the SAS software.

The results obtained from the DA model were analyzed, and the confusion matrix for the statistical technique is presented in **Table 9**, also obtained from SAS.

This additional analysis using DA provides further insights into the relationship between the independent variables and the dependent variable (diabetes) in our dataset [11]. By comparing the results of the DA model to those of the LR model, we can better understand the robustness and generalizability of our findings.

The confusion table (refer to **Table 9**) shows that there is a total of 390 patients (330 of which are patients without diabetes while 60 are patients with diabetes). Of the 330 patients without diabetes, 310 patients were correctly predicted as non-

diabetic while 20 were wrongly classified as diabetic), we can state that the model gave correct prediction for approximately 94% of the patients without diabetes. Additionally, the model wrongly predicted 11 diabetes patients as "non-diabetic" and appropriately predicted 49 patients as actually having diabetes. We can say that the model prediction here is around 82%.

Table 9. Classification summary for discriminant analysis.

Number of Observations and Percent Classified into Diabetes			
From Diabetes	**0**	**1**	**Total**
0	310 93.94	20 6.06	330 100.00
1	11 18.33	49 81.67	60 100.00
Total	321 82.31	69 17.69	390 100.00
Priors	0.5	0.5	

The overall accuracy of a binary classification is a measure of how often the model correctly predicts the class of new observations. It is calculated as follows:

$$\text{Overall accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}},$$

where TP = True Positive, TN = True Negative, FP = False Positive and FN = False Negative.

The overall prediction of the DA model can therefore be computed as follows:

TN = 310, TP = 49, FN = 20 and FP = 11. The prior probabilities were 50:50 (an equally likely chance scenario)

$$\frac{310 + 49}{310 + 49 + 20 + 11} \times 100 = 92\%$$

That means that the DA model correctly predicted outcomes (class labels) for 92% of the sample population provided.

Table 10. Performance metrics.

Performance Metric	Score
Overall accuracy	92%
Specificity	97%
Sensitivity	71%
Negative Predictive Value	94%
Positive Predictive Value	82%

Performance metrics were computed from the confusion matrix presented in Table 10, using the following standard formulas: Metrics Formula Value Overall

Accuracy (TP + TN)/(TP + TN + FP + FN) 92.05%, Sensitivity (Recall) TP/(TP + FN) 71.01%, Specificity TN/(TN + FP) 96.57%, Precision (PPV) TP/(TP + FP) 81.67%, Negative Predictive Value (NPV) TN/(TN + FN) 93.94%.

The high overall accuracy and specificity suggests that the DA model effectively identifies non-diabetic individuals [10]. At the same time, the lower sensitivity implies a higher chance of false negatives in identifying diabetic individuals. The high negative predictive value means that the model effectively identifies individuals who do not have diabetes. In contrast, the positive predictive value indicates that the model can correctly identify 8 out of every 10 individuals who have diabetes.

4.5. Principal Component Analysis

The dataset was subjected to Principal Component Analysis (PCA) to reduce the number of attributes to the optimal number of variables that capture the variability in the data [19]. From the analysis done in SAS, it was observed that five principal components are sufficient to run the model, as their eigenvalues were greater than 1. The results indicate that the Logistics Regression and Discriminant Analysis procedures can be run using these five principal components, which explain approximately 78.1% of the variance in the dataset. Table 11 displays the eigenvalues and the proportion of variance explained by each principal component.

Table 11. Eigenvalues of the correlation matrix.

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	4.08527421	2.06260964	0.3143	0.3143
2	2.02266457	0.33165910	0.1556	0.4698
3	1.69100547	0.39560790	0.1301	0.5999
4	1.29539756	0.23773573	0.0996	0.6996
5	1.05766184	0.09126156	0.0814	0.7809
6	0.96640027	0.14253078	0.0743	0.8553
7	0.82386949	0.25632787	0.0634	0.9186
8	0.56754162	0.29866870	0.0437	0.9623
9	0.26887292	0.12323370	0.0207	0.9830
10	0.14563921	0.07814038	0.0112	0.9942
11	0.06749884	0.06132666	0.0052	0.9994
12	0.00617218	0.00417034	0.0005	0.9998
13	0.00200184		0.0002	1.0000

The scree plot elbow also establishes that five principal components are statistically sufficient to explain the variance (please see Figure 5).

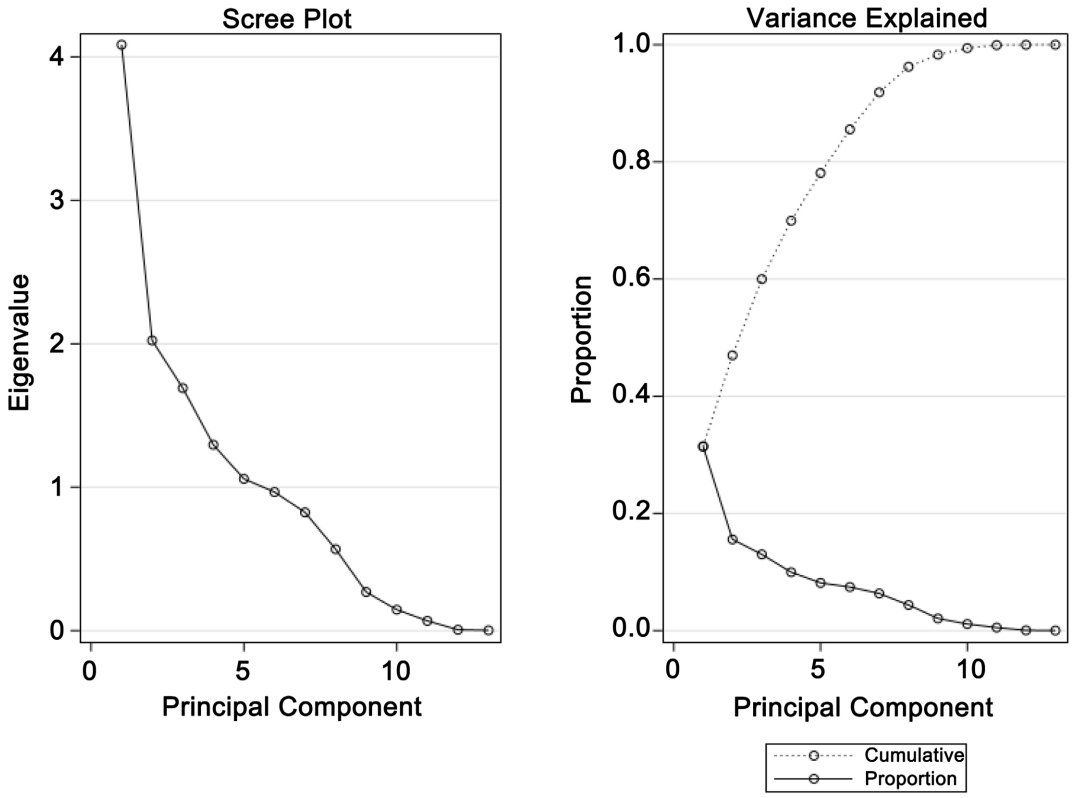

Figure 5. Scree plot and variance explanation.

4.6. Logistic Regression Results (after PCA)

We fitted Logistic Regression to the PCA-reduced dataset. We checked the output file for the final results and the goodness of fit by inspecting the -2LogL values, the AIC (Akaike Information Criterion), SC (Schwarz's Bayesian Criterion). We checked the difference in the values as the intercepts, prin1, prin2, prin3, prin4 and prin5 using *the stepwise procedure* (*forward selection*). **Figure 6** presents screenshots results of the step by step process utilized. The p-values (≤0.001) are less than 0.05 (the significance level), indicating that the model is statistically significant at every level of variable addition. The result section is given below:

Step 0. Intercept entered:

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

-2 Log L		334.872

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
112.8858	5	<.0001

(a)

Step 1. Effect Prin2 entered:

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	336.872	295.674
SC	340.838	303.607
-2 Log L	334.872	291.674

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	43.1976	1	<.0001
Score	43.4239	1	<.0001
Wald	36.8933	1	<.0001

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
65.5955	4	<.0001

Note: No effects for the model in Step 1 are removed.

(b)

Step 2. Effect Prin1 entered:

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	336.872	257.090
SC	340.838	268.989
-2 Log L	334.872	251.090

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	83.7818	2	<.0001
Score	81.3143	2	<.0001
Wald	56.3456	2	<.0001

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
29.2196	3	<.0001

Note: No effects for the model in Step 2 are removed.

(c)

Step 3. Effect Prin5 entered:

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	336.872	240.513
SC	340.838	256.378
-2 Log L	334.872	232.513

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	102.3586	3	<.0001
Score	98.7210	3	<.0001
Wald	65.4034	3	<.0001

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
12.5293	2	0.0019

Note: No effects for the model in Step 3 are removed.

(d)

Step 4. Effect Prin3 entered:

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	336.872	234.379
SC	340.838	254.210
-2 Log L	334.872	224.379

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	110.4931	4	<.0001
Score	108.2813	4	<.0001
Wald	66.6331	4	<.0001

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
4.9606	1	0.0259

Note: No effects for the model in Step 4 are removed.

(e)

Step 5. Effect Prin4 entered:

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	336.872	231.358
SC	340.838	255.155
-2 Log L	334.872	219.358

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	115.5136	5	<.0001
Score	112.8858	5	<.0001
Wald	66.3298	5	<.0001

Note: No effects for the model in Step 5 are removed.
Note: All effects have been entered into the model.

(f)

Figure 6. (a-f) Screen shot of model convergence status.

Table 12. Validating model fit using principal component analysis (PCA).

Procedure	0	1	2	3	4	5
AIC		336.872	336.872	336.872	336.872	336.872
SC		340.838	340.838	340.838	340.838	340.838
−2logL	334.872	334.872	334.872	334.872	334.872	334.872

As seen in Figures 6(a)-(f) and Table 12, the values of the parameters (AIC, SC, and −2logL) remained unchanged. Theoretically, this implies that the addition of new variables (principal components) did not directly improve the model, but this is not to say that the principal components are not outrightly irrelevant.

The Logistic Regression model below is formulated according to the maximum likelihood estimates obtained from SAS and presented in Table 13.

Table 13. Maximum likelihood.

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	−2.5661	0.2432	111.2905	<0.0001
Prin1	1	0.5689	0.099	32.4574	<0.0001
Prin2	1	0.7415	0.1314	31.8303	<0.0001
Prin3	1	−0.3527	0.1235	8.1559	0.0043
Prin4	1	−0.3473	0.1578	4.8446	0.0277
Prin5	1	0.6877	0.1612	18.1996	<0.0001

$$\ln\left(\frac{p}{1-p}\right) = -2.5661 + 0.5689 \times \text{Prin1} + 0.7415 \times \text{Prin2} - 0.3527 \times \text{Prin3} - 0.3473 \times \text{Prin4} + 0.6877 \times \text{Prin5}$$

4.7. Discriminant Analysis (with PCA)

Using the PCA-reduced dataset, the DA model produced classification results (Table 14), and two discriminant functions (Table 15) which were derived in SAS.

Table 14. Classification summary for discriminant analysis with PCA.

Number of Observations and Percent Classified into Diabetes			
From Diabetes	**0**	**1**	**Total**
0	270 81.82	60 18.18	330 100.00
1	15 25.00	45 75.00	60 100.00
Total	285 73.08	105 26.92	390 100.00
Priors	0.5	0.5	

Table 15. Linear discriminant function.

Linear Discriminant Function for Diabetes		
Variable	**0**	**1**
Constant	−0.03684	−1.11450
Prin1	−0.09219	0.50703
Prin2	−0.14026	0.77140
Prin3	0.07197	−0.39586
Prin4	0.05707	−0.31388
Prin5	−0.12280	0.67540

$$D_0(\text{No Diabetes}) = -0.03684 - 0.09219 \times \text{Prin1} - 0.14026 \times \text{Prin2} + 0.07197 \times \text{Prin3} + 0.05707 \times \text{Prin4} - 0.12280 \times \text{Prin5}$$

$$D_1(\text{Diabetes present}) = -1.11450 + 0.50703 \times \text{Prin1} + 0.77140 \times \text{Prin2} - 0.39586 \times \text{Prin3} - 0.31388 \times \text{Prin4} + 0.67540 \times \text{Prin5}$$

4.8. Comparing both Statistical Techniques

4.8.1. Logistic Regression (Before and After PCA Transformation)

The values of the goodness of fit (−2logL, AIC, and SC) were examined for the LR model outcomes on the original diabetes dataset and after the PCA transformation. The results, presented in Table 16, showed that these values remained unchanged for both procedures, indicating that PCA did not enhance the model's performance. PCA is typically used to improve model accuracy in datasets with highly correlated independent variables. However, in this study, PCA did not sig-

nificantly affect the model performance, likely because only a few variables had high correlation values. Additionally, comparing the AIC and SC values before and after PCA transformation, which were 336.872 and 340.838, respectively, further confirmed that the procedure did not improve the model fit. In general, to achieve a better model fit, the values of −2logL, AIC, and SC should decrease after the PCA transformation, with smaller values indicating better performance.

Table 16. Goodness of fit variables.

	Before PCA	After PCA
−2logL	334.872	334.872
AIC	336.872	336.872
SC	340.878	340.878

4.8.2. Discriminant Analysis (Before and After PCA Transformation)

Table 17 shows the comparison of the results obtained for the DA procedure before the PCA was done and after. It was observed that the overall model accuracy reduced from 92% to 81% including specificity and sensitivity.

Table 17. Before and after DA.

Performance Metric	Before PCA	After PCA	
Overall Accuracy	92%	81%	−11%
Specificity	97%	95%	−2%
Sensitivity	71%	43%	−28%
Negative Predictive Value	94%	82%	−12%
Positive Predictive Value	82%	75%	−7%

4.8.3. Cross-validation of LR and DA on the Original Dataset

In order to compare the accuracy of the Logistic Regression and Discriminant Analysis and check which procedure is better, cross-validation procedure was

```
In [28]: X = data.drop("Diabetes",axis=1)
y = data['Diabetes']
from sklearn.model_selection import KFold, cross_val_score
# Create a linear regression model

X_train, X_test, y_train, Y_test = train_test_split(X, y, test_size=0.2, random_state= 42)
# Define the number of folds for cross-validation
k_folds = KFold(n_splits=5)
model =LogisticRegression(max_iter=2000)
# Perform cross-validation and compute the mean and standard deviation of the scores
scores = cross_val_score(model, X, y, cv=k_folds)
mean_score = np.mean(scores)
std_score = np.std(scores)
# Print the results
print("Cross-validation scores: ", scores)
print("Mean score: ", mean_score)
print("Standard deviation: ", std_score)

Cross-validation scores:  [0.98591549 0.98591549 0.94366197 0.91549296 0.8028169 ]
Mean score:  0.9267605633802818
Standard deviation:  0.0674881608490757
```

Figure 7. Snapshot of cross validation result for Logistic Regression.

```
In [30]: X = data.drop("Diabetes",axis=1)
         y = data['Diabetes']
         from sklearn.model_selection import KFold, cross_val_score
         # Create a linear discriminant analysis model

         X_train, X_test, y_train, Y_test = train_test_split(X, y, test_size=0.2, random_state= 42)

         # Define the number of folds for cross-validation
         k_folds = KFold(n_splits=4)
         model =LinearDiscriminantAnalysis()
         # Perform cross-validation and compute the mean and standard deviation of the scores
         scores = cross_val_score(model, X, y, cv=5)
         mean_score = np.mean(scores)
         std_score = np.std(scores)
         # Print the results
         print("Cross-validation scores: ", scores)
         print("Mean score: ", mean_score)
         print("Standard deviation: ", std_score)

         Cross-validation scores:  [0.92957746 0.95774648 0.91549296 0.94366197 0.87323944]
         Mean score:  0.923943661971831
         Standard deviation:  0.029001775045033804
```

Figure 8. Snapshot of cross validation result for discriminant analysis.

done on the original diabetes dataset using Python Integrated Development Environment (code snippets for this procedure are presented in **Figure 7** and **Figure 8** respectively). From the sklearn.model selection library, Kfold and cross_val_score are imported.

The result obtained for the k-fold cross validation shows a mean score of 92.67% for the LR model and 92.39% for the DA model. Both statistical techniques gave very good results.

4.9. Implications for Clinical Practice

The hazard risk for diabetes mellitus is rising [20]. Accurate prediction analysis can significantly benefit clinical practice. Other studies have indicated large waist circumference, high HgbA1C, and high Fatty Liver Index (FLI) all indicate the greater possibility of contracting diabetes [21]. Our dataset has one similar variable, the waist-to-hip ratio and others like glucose and BMI. Our list of variables may be more reflective of collection site limited data. Regardless, we are able to provide a valid prediction model.

Uncontrolled glucose levels in the bloodstream can increase damage to the body. Serious damage to the eyes, kidneys, and vasculature (heart) require extensive medical intervention. Predicting the likelihood of becoming a diabetic can help direct educational resources to those most in need of reversing this trend. Education focuses on a healthy diet and frequent exercise to improve long-term health outcomes [22]. Computational intensity of neural networks and the need for large datasets, limit complex model usefulness in a clinical practice environment. More likely, these techniques can be found in data analyst environments. Given the small size of our dataset, LR and DA adequately supported prediction analysis without extensive time and costs. Having a data analyst on the healthcare team could be beneficial.

4.10. Limitations of LR and DA

There are at least four critical assumptions to consider when choosing LR and DA. We rely on assumptions about the data. Violating these assumptions can lead to inaccurate results. The assumptions are:

1) Linearity of data
2) Normality of residuals
3) Homogeneity of residuals variance
4) Independence of residuals error terms

Before applying a technique, it is crucial to ensure that the assumptions are met. Regression beta coefficients and R2 can be used to tell how well the LR model fits to the data.

There are potential problems to address:

1) Non-linearity
2) Heteroscedasticity
3) Presence of influential values

These problems can be checked with diagnostic plots in R software. Moderate multicollinearity may not be problematic. However, severe multicollinearity is a problem because it can increase the variance of the coefficient estimates and make the estimates very sensitive to minor changes in the model. The result is that the coefficient estimates are unstable and difficult to interpret. Multicollinearity saps the statistical power of the analysis, can cause the coefficients to switch signs, and makes it more difficult to specify the correct model (Regression Analysis blog). High predictor correlation or multicollinearity can impact the interpretability of the regression model. However, a reasonable prediction can still be made, if attention is directed toward eliminating inflated standard errors and preventing overfitting.

5. Conclusions

In this study, Principal Component Analysis (PCA) was applied as an exploratory step for training the Logistic Regression (LR) and Discriminant Analysis (DA) models. The goal was to observe whether reducing the number of input features in the form of principal components could improve model performance. However, no improvement was observed after PCA was applied. This outcome can be explained by the nature of both the dataset and the models used. The correlation matrix of the original Diabetes dataset did not reveal strongly correlated features, which reduced the need for dimensionality reduction. Although some high correlation was present between Waist and Weight (0.8478), Hip and Weight (0.8270), and Waist and Hip (0.8352), the majority of variables did not exhibit strong collinearity. Furthermore, both LR and DA are linear models and are capable of handling moderately correlated features. Applying a linear transformation such as PCA did not introduce any advantage and may have discarded useful information. As a result, PCA was not used further for developing and tuning the final models. Post-PCA correlation structure was not computed, as PCA by design generates uncorrelated principal components. Since PCA did not improve model perfor-

mance, and the original dataset showed limited strong correlations, further correlation analysis post-PCA was not pursued. For the exploratory analysis, an optimal number of principal components were selected based on the scree plot presented in **Figure 4**. The values of the principal components are the loading coefficients obtained from the Maximum Likelihood Estimates and Linear Discriminant Analysis procedures performed using SAS and are presented in **Table 7** and **Table 9**.

The effectiveness of Logistic Regression and Discriminant Analysis techniques in accurately predicting diabetes in individuals was demonstrated. The results obtained from our Diabetes dataset show that the use of PCA did not enhance the performance of these models. This reflects a weak relationship between variables. According to Praxis Business School, PCA is ineffectual when the correlation matrix is below 0.3 as is the case for most of the correlation coefficients that were observed. However, it is important to note that our study is limited by the relatively small size of our dataset, which may restrict the generalization of our findings.

To build on these results, future research can focus on training these models to adapt to larger datasets, as this may provide more robust and reliable results. Additionally, the use of more advanced intelligent systems, such as Artificial Neural Networks (ANN) [13] and ensemble models, can be explored to further improve the accuracy of predicting diabetes in individuals [12].

Furthermore, it would be interesting to investigate the performance of these models based on gender demographic. This could provide useful insights into how diabetes manifests differently in males and females and could lead to more tailored and effective interventions for diabetes prevention and management.

Overall, while our study has some limitations, our findings underscore the usefulness of Logistic Regression and Discriminant Analysis techniques in predicting diabetes and highlight the need for further research to continue to improve the accuracy and effectiveness of these models.

Author Contributions

E.D., F.I. and M.O. contributed equally to the conceptualization and methodology of this original research. All authors have read and agreed to the published version of the manuscript.

Funding

The authors would like to express their gratitude for the funding support from the Department of Industrial and Systems Engineering, North Carolina Agricultural and Technical State University.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] Centers for Disease Control (2024) CDC Diabetes Basics.

https://www.cdc.gov/diabetes/about/index.html

[2] Gwira, J.A., Fryar, C.D. and Gu, Q.P. (2024) Prevalence of Total, Diagnosed, and Undiagnosed Diabetes in Adults: United States, August 2021-August 2023. NCHS Data Briefs, National Center for Health Statistics (US).

[3] Centers for Disease Control (2024) National Diabetes Statistics Report. https://www.cdc.gov/diabetes/php/data-research/methods.html?CDC_AAref_Val=https://www.cdc.gov/diabetes/data/statistics-report/index.html?ACSTrackingID=DM72996&ACSTrackingLabel=New%2520Report%2520Shares%2520Latest%2520Diabetes%2520Stats%2520&deliveryName=DM72996

[4] Walker, R.J., Garacci, E., Ozieh, M. and Egede, L.E. (2021) Food Insecurity and Glycemic Control in Individuals with Diagnosed and Undiagnosed Diabetes in the United States. *Primary Care Diabetes*, **15**, 813-818. https://doi.org/10.1016/j.pcd.2021.05.003

[5] Berkowitz, S.A., Karter, A.J., Corbie-Smith, G., Seligman, H.K., Ackroyd, S.A., Barnard, L.S., *et al.* (2018) Food Insecurity, Food "Deserts," and Glycemic Control in Patients with Diabetes: A Longitudinal Analysis. *Diabetes Care*, **41**, 1188-1195. https://doi.org/10.2337/dc17-1981

[6] Abdulhadi, N. and Al-Mousa, A. (2021) Diabetes Detection Using Machine Learning Classification Methods. 2021 *International Conference on Information Technology* (*ICIT*), Amman, 14-15 July 2021, 350-354. https://doi.org/10.1109/icit52682.2021.9491788

[7] Zhu, C., Idemudia, C.U. and Feng, W. (2019) Improved Logistic Regression Model for Diabetes Prediction by Integrating PCA and K-Means Techniques. *Informatics in Medicine Unlocked*, **17**, Article 100179. https://doi.org/10.1016/j.imu.2019.100179

[8] Rajendra, P. and Latifi, S. (2021) Prediction of Diabetes Using Logistic Regression and Ensemble Techniques. *Computer Methods and Programs in Biomedicine Update*, **1**, Article 100032. https://doi.org/10.1016/j.cmpbup.2021.100032

[9] Joshi, R.D. and Dhakal, C.K. (2021) Predicting Type 2 Diabetes Using Logistic Regression and Machine Learning Approaches. *International Journal of Environmental Research and Public Health*, **18**, Article 7346. https://doi.org/10.3390/ijerph18147346

[10] Polat, K., Güneş, S. and Arslan, A. (2008) A Cascade Learning System for Classification of Diabetes Disease: Generalized Discriminant Analysis and Least Square Support Vector Machine. *Expert Systems with Applications*, **34**, 482-487. https://doi.org/10.1016/j.eswa.2006.09.012

[11] Alharan, A.F.H., Algelal, Z.M., Ali, N.S. and Al-Garaawi, N. (2021) Improving Classification Performance for Diabetes with Linear Discriminant Analysis and Genetic Algorithm. 2021 *Palestinian International Conference on Information and Communication Technology* (*PICICT*), Gaza, 28-29 September 2021, 38-44. https://doi.org/10.1109/picict53635.2021.00019

[12] El Jerjawi, N.S. and Abu-Naser, S.S. (2018) Diabetes Prediction Using Artificial Neural Network. *International Journal of Advanced Science and Technology*, **121**, 55-64.

[13] Srivastava, S., Sharma, L., Sharma, V., Kumar, A. and Darbari, H. (2018) Prediction of Diabetes Using Artificial Neural Network Approach. In: Ray, K., Sharan, S., Rawat, S., Jain, S., Srivastava, S. and Bandyopadhyay, A., Eds., *Lecture Notes in Electrical Engineering*, Springer, 679-687. https://doi.org/10.1007/978-981-13-1642-5_59

[14] Tabaei, B.P. and Herman, W.H. (2002) A Multivariate Logistic Regression Equation to Screen for Diabetes. *Diabetes Care*, **25**, 1999-2003.

https://doi.org/10.2337/diacare.25.11.1999
http://diabetesjournals.org/care/article-pdf/25/11/1999/588640/dc1102001999.pdf

[15] Ye, Y., Xiong, Y., Zhou, Q., Wu, J., Li, X. and Xiao, X. (2020) Comparison of Machine Learning Methods and Conventional Logistic Regressions for Predicting Gestational Diabetes Using Routine Clinical Data: A Retrospective Cohort Study. *Journal of Diabetes Research*, **2020**, Article ID: 4168340. https://doi.org/10.1155/2020/4168340

[16] Vihinen, M. (2012) How to Evaluate Performance of Prediction Methods? Measures and Their Interpretation in Variation Effect Analysis. *BMC Genomics*, **13**, S2. https://doi.org/10.1186/1471-2164-13-s4-s2

[17] Akobeng, A.K. (2007) Understanding Diagnostic Tests 1: Sensitivity, Specificity and Predictive Values. *Acta Paediatrica*, **96**, 338-341. https://doi.org/10.1111/j.1651-2227.2006.00180.x

[18] Nti, I.K., Nyarko-Boateng, O. and Aning, J. (2021) Performance of Machine Learning Algorithms with Different K Values in K-Fold Cross-Validation. *International Journal of Information Technology and Computer Science*, **13**, 61-71.

[19] Rahimloo, P. and Jafarian, A. (2016) Prediction of Diabetes by Using Artificial Neural Network, Logistic Regression Statistical Model and Combination of Them. *Bulletin de la Société Royale des Sciences de Liège*, **85**, 1148-1164.

[20] Hounguè, P. and Bigirimana, A.G. (2022) Leveraging Pima Dataset to Diabetes Prediction: Case Study of Deep Neural Network. *Journal of Computer and Communications*, **10**, 15-28. https://doi.org/10.4236/jcc.2022.1011002

[21] Tanaka, M., Akiyama, Y., Mori, K., Hosaka, I., Kato, K., Endo, K., *et al.* (2024) Predictive Modeling for the Development of Diabetes Mellitus Using Key Factors in Various Machine Learning Approaches. *Diabetes Epidemiology and Management*, **13**, Article 100191. https://doi.org/10.1016/j.deman.2023.100191

[22] Feinberg, A., *et al.* (2018) Prescribing Food as a Specialty Drug. *NEJM Catalyst*.