Prediction of Type 2 Diabetes using logistic regression techniques

Ghadeer Mousa, Hassan Abu Hassan and Hussein Al-Rimmawi
Birzeit University, Palestine

Abstract: Importance of the Problem: Diabetes is recognized as a significant public health concern and a global epidemic. It is a chronic condition resulting from insufficient insulin production by the pancreas. The long-term elevated blood sugar levels associated with diabetes lead to chronic damage and impaired function in multiple tissues, such as the eyes, kidneys, heart, blood vessels, and nerves.

The objective of this study is to demonstrate the utilization of machine-learning algorithms, specifically logistic regression, in predicting an individual's likelihood of having diabetes based on medical data. Furthermore, the study aims to develop a prediction model that determines whether a patient has diabetes by analyzing specific diagnostic measurements included in the dataset. Various techniques will be explored to enhance the performance and accuracy of the prediction model.

Results: The logistic regression algorithm for the dataset containing various patient data, found that the algorithm predicted whether people would be diagnosed with diabetes with an 82 percent success rate**.**

Keywords: Machine Learning, Type 2 Diabetes, Logistic Regression (Times New Roman 9)

1. Introduction (Times New Roman 10 Bold)

Diabetes mellitus refers to a collection of metabolic disorders characterized by elevated blood sugar levels (hyperglycemia) resulting from deficiencies in insulin action, insulin secretion, or both (American Diabetes Association, 2009, 1-8). The International Diabetes Federation (IDF) reported that in 2017, there were 425 million individuals worldwide living with diabetes. However, by 2019, this number had risen to 463 million adults aged 20 to 79 years, highlighting the alarming increase and positioning diabetes as a significant global health crisis in the 21st century (Rajendraand & Latifi, 2021, 1-8).

Types of Diabetes:

Diabetes is classified into three distinct types: type 1 diabetes, type 2 diabetes, and gestational diabetes. Gestational Diabetes: This form of diabetes emerges during pregnancy and may potentially resolve after childbirth. However, if left untreated, it carries a heightened risk of progressing into type 2 diabetes. Type 1 Diabetes: This type arises when the body produces insufficient or no insulin at all. It predominantly affects children, teenagers, and young adults, and is characterized by a deficiency of insulin. Individuals with type 1 diabetes require insulin injections for management. The precise cause of this type of diabetes remains unknown. Symptoms encompass increased urination (polyuria), excessive thirst (polydipsia), constant hunger, weight loss, vision changes, and fatigue. These symptoms may manifest suddenly (Cho et al., 2018, 271-281).

Type 2 Diabetes:

Insulin resistance is the underlying cause of this type. While it predominantly affects adults, there is a growing prevalence of type 2 diabetes among children as well. Individuals with type 2 diabetes have insufficient levels of insulin in their bodies. It accounts for over 95% of all diabetes cases. The primary factors contributing to type 2 diabetes are excess body weight and a sedentary lifestyle. Although the symptoms resemble those of type 1 diabetes, they are generally less severe. Consequently, the diagnosis of this condition often occurs years later, after complications have already developed (World Health Organization, 2019).

Factors Responsible for Diabetes

Diabetes can be attributed to a combination of genetic susceptibility and environmental factors. Prolonged overweight can lead to diabetes, as well as being born into a family with a history of the condition. Furthermore, the risk of developing type 2 diabetes tends to increase gradually as we age. Several potential complications are associated with diabetes, including cardiovascular diseases such as high blood pressure, heart failure, stroke, and even death resulting from heart attacks or other conditions related to arterial hypertension (elevated pressure within the arteries) (Murea, Lijun & Freedman, 2012, 6-22).

Machine learning:

Machine learning, which falls under the umbrella of artificial intelligence, addresses practical challenges by empowering computers to learn autonomously without explicit programming (Dhage & Raina, 2016, 395-399). In the realm of disease detection and diagnosis, machine learning systems are often designed to emulate the expertise of medical professionals. By leveraging data, machine learning algorithms can develop models that identify recurring patterns and make informed decisions, thereby mitigating shortcomings in the medical databases utilized (Dhage & Raina, 2016, 395-399).

Literature Review:

In the literature, researchers have employed diverse data mining approaches and machine learning algorithms to propose and implement numerous prediction methods. In the past decade, there has been a notable emphasis on the creation of predictive models specifically tailored for diabetes.

In their work, Orabi, Kamal, and Rabah (2016, 420-427) presented an early predictive system for Diabetes Mellitus Disease, achieving an accuracy rate of 84%. The system is capable of predicting the likelihood of an individual being a candidate for diabetes and estimating the age at which it may occur. The study utilized an Egyptian diabetic dataset, with two-thirds of the data used for training the model and one-third for testing its performance.

In a study conducted by Alajlan (2021, 3957-3965), three machine learning algorithms, namely decision tree, AdaBoost, and KNN, were investigated for the purpose of predicting early-stage diabetes. The evaluation of these algorithms involved the use of sensitivity, precision, f-measure, and ROC curve as performance metrics. Among the algorithms tested, Adaptive Boosting (AdaBoost) demonstrated the highest performance and outperformed many other supervised machine learning classification algorithms in terms of predictive accuracy.

In their study, Lai et al. (2019, 1-9) employed Logistic Regression and Gradient Boosting Machine (GBM) techniques to construct predictive models for diabetes mellitus. The study utilized data from 13,309 Canadian patients aged between 18 and 90 years old. The variables considered in the study included age, sex, fasting blood glucose, body mass index, high-density lipoprotein, triglycerides, blood pressure, and low-density lipoprotein. The findings of the study highlighted that fasting blood glucose, body mass index, high-density lipoprotein, and triglycerides emerged as the most crucial predictors in the models for identifying patients who may develop diabetes in the future. This information can be instrumental in implementing necessary preventive interventions. In a study conducted in Gaza, Palestine, by Al-Saftawi, Harz, and Hijazi (2021, 61-70), an artificial neural network was employed to predict whether an individual is diabetic or not. The study involved a sample of 520 individuals and 17 attributes. The neural network model was trained to minimize the error function during the training process. Upon training the model, it achieved an average error function of 0.01, indicating a high level of accuracy. The prediction accuracy for determining whether a person is diabetic or not was reported to be 98.84%. In their study, Joshi and Dhakal (2021, pp. 1-7) employed logistic regression, a widely used classification methodology. The study focused on identifying the impact of five variables, namely age, body mass index, pedigree, glucose, and frequency of regencies, on disease prevention. By controlling these variables, the researchers aimed to design effective interventions and implement health policies to prevent the onset of the disease.

In their study, Rahimloo and Jafarian (2016, pp. 1148-1164) utilized a hybrid neural network model that combined artificial neural networks with logistic regression to predict diabetes. The primary objective of their research was to identify the significant variables and their impact on diabetes prediction. By comparing different methods, the researchers concluded that the prediction of diabetes can be improved by employing the method with the lowest error rate, thereby enhancing the accuracy of the predictions.

In their study, Daghistani and Alshammari (2016, 329-332) conducted a comparison between the Random Forest machine learning algorithm and the Logistic Regression algorithm for the prediction of diabetes. The study utilized a dataset comprising 66,325 records from a Saudi Arabian hospital spanning the years 2013 to 2015. The dataset encompassed 18 different risk factors. The researchers discovered that the Random Forest model exhibited superior predictive performance compared to the Logistic Regression model.

In their study, Cahyani et al. (2022, 107-114) aimed to predict the presence of diabetes in patients. They utilized the logistic regression algorithm with normalization in their analysis. The researchers concluded that by employing normalization techniques, the diabetes risk prediction improved to 55%, whereas without normalization it yielded a prediction accuracy of 43%.

In the Middle East, Bhar and Abu Sadah (2018, 231-261) utilized logistic regression to develop a statistical model for classifying diabetes data. The study sample included 232 individuals, comprising 172 with diabetes and 60 without diabetes. The sample encompassed individuals of both genders aged 25 years and above. Data were collected from public and private health clinics in the Gaza Strip, Palestine. The developed model achieved a precision rate of 92%. The study identified several factors that influenced diabetes, including psychological pressure, weekly consumption of fruits and meat, and the level of education. To evaluate the model, techniques such as out-one-leave cross-validation, classification tables, and ROC curves were employed.

In the field of heart diseases, Abu Duma (2019) employed both Logistic Regression and Discriminant Analysis techniques to identify factors influencing heart disease infection. The study specifically focused on a comparative analysis between the Cardiac Surgery and Renal Transplant Center at Ahmed Gasim Hospital in Khartoum North. The sample size consisted of 214 individuals who were infected with heart diseases and 214 non-infected individuals. The research revealed that hypertension, gender, diabetes, cholesterol levels, hereditary factors, and weight were significant factors affecting heart disease. The predictive precision of the model reached 92%. Abu Duma also determined that the binary logistic regression model outperformed the discriminant model due to its higher efficiency and its ability to diagnose and predict the likelihood of infection with a minimal error rate of 8.2%. Furthermore, the study recommended the inclusion of additional variables such as age, smoking, and physical exercise in future investigations.

Khalel (2016, 220-246) utilized descriptive statistics and logistic regression to investigate the factors contributing to marriage delay among the population of Umluj University in Saudi Arabia. The study aimed to identify the reasons behind this delay. Several variables were found to have an impact on marriage delay, including the loss of a parent, high expectations regarding future partners, paternal dominance over children, individuals' perceptions of the families they intend to form, high housing rents, low income, and the pursuit of education. The use of descriptive statistics and logistic regression allowed for a comprehensive analysis of these factors and their influence on marriage delay in the studied population.

Khalel (2016, 220-246) employed descriptive statistics and logistic regression in order to find out reasons which causes marriage delay among population of Umluj University, Saudi Arabia. Several variables affect marriage delay. These variables are loss of parent, high expectation of future partner, Father’s domination over children, individuals estimate of the family he intends to form, high housing rents, low income and pursuing education.

Types of Machine learning:

Machine learning techniques are widely used for dataset classification, employing approaches such as supervised learning, unsupervised learning, reinforcement learning, and deep learning.

Supervised learning involves training algorithms using labeled examples, where the input data is paired with corresponding target outputs. By learning from this training set, the algorithms can accurately predict outputs for new, unseen inputs. Supervised learning includes two main categories: classification, which predicts discrete classes or categories such as "Yes" or "No," and regression, which estimates continuous values like "How much" or "How many."

Unsupervised learning, on the other hand, aims to find patterns or structures within unlabeled data. It does not have predefined target outputs but instead focuses on discovering relationships or similarities between data points. Common techniques in unsupervised learning include clustering, where data is grouped based on similarities or proximity, and density estimation, which estimates the underlying probability distribution of the data.

Reinforcement learning is a different paradigm where an agent learns to interact with an environment through trial and error. It receives feedback in the form of rewards or penalties based on its actions and aims to maximize its cumulative reward over time. Unlike supervised learning, reinforcement learning does not rely on labeled examples but rather learns through exploration and exploitation of the environment.

Deep learning is a subset of machine learning that utilizes neural networks with multiple layers to extract highlevel features and representations from data. These deep neural networks can model complex relationships and patterns in the data, and they excel at tasks such as image and speech recognition. Deep learning has achieved remarkable success in various domains, thanks to its ability to automatically learn hierarchical representations from large amounts of data.

Overall, these different machine learning techniques offer powerful tools for classifying datasets and solving a wide range of real-world problems.

The goal of supervised learning algorithms is indeed to predict labeled data. In supervised learning, the algorithm is provided with a training dataset that consists of input features and corresponding labeled outputs. The algorithm learns from this labeled data to make predictions or classifications on new, unseen data.(Fatima & Maruf, 2017, 116).

Logistic Regression: Logistic regression, a widely recognized technique adopted from statistics in the field of machine learning, utilizes real-valued inputs to estimate the probability of an input belonging to a specific class, such as the diabetes class (referred to as class 0). When the predicted probability exceeds 0.5, it classifies the input as class 0; otherwise, it is classified as class 1. This classification algorithm employs one or more independent features to determine the outcome. Given that our dependent variable, 'Outcome,' has only binary values (0 and 1), logistic regression was the most straightforward approach for training the dataset (Brownlee, 2020).

Research Problem: Diabetes diagnosis is critical for active care in persons who are newly diagnosed and have not yet acquired complications. Such people did not have the chance in advance to be aware of the early diabetes symptoms. It is unrealistic to expect everyone to be aware of the early symptoms. Therefore, this research focuses on a potential system that can assist a healthcare practitioner to early detect of diabetes using one of the frequently utilized classification algorithms.

Research Objectives: The objectives of this research are:
To address the classification of Diabetes Mellitus using a logistic regression classifier, our objective is to apply and support the implementation of the logistic regression classification technique. This will aid in standardizing the diagnosis of Diabetes Mellitus within the dataset of patients.

Research Methodology:
Description of Pima Indigenous Dataset: In the text, the authors used the term "indigenous" instead of "Indian." The dataset utilized in this project is known as the Pima Indigenous Diabetes database, which was sponsored and published by the National Institute of Diabetes, Digestive and Kidney Diseases in the United States of America. It is publicly accessible on the Kaggle website (https://www.kaggle.com/uciml/pima-indians-diabetes-database) and serves as an open-source dataset comprising records of female patients. The dataset encompasses a total of 768 cases, with each case representing a female participant from the Pima Indigenous community (table 1). Within each case, there is a binary indicator indicating whether the individual is non-diabetic (0) or diabetic (1). The dataset contains 500 cases classified as non-diabetic and 268 cases classified as diabetic. Additionally, the dataset includes the following eight features."
Pregnancies: Number of times a Pima Indigenous female got pregnant.

Glucose Level: Plasma glucose concentration over 2 hours in an oral glucose tolerance test.

Blood pressure: Blood pressure refers to the force exerted by the blood as it circulates through the body's cardiovascular system. It plays a crucial role in maintaining proper circulation. Both high and low blood pressures can have significant implications for one's health, and extreme fluctuations in blood pressure can even serve as an indicator of potential mortality.

Skin Thickness: Triceps skinfold thickness, measured in millimeters (mm) within the dataset, is a metric that offers a reliable estimate of both obesity and body fat distribution. It serves as a valuable indicator in assessing body composition and provides insights into the distribution of fat in the triceps region of the body.

Insulin:
In the Pima Indigenous community, the metric used to measure the level of insulin in the blood after a two-hour period is denoted as "mu U/ml" (micro-units per milliliter). In the context of the dataset, the variable labeled 'Insulin' represents the two-hour serum insulin level. By analyzing an individual's insulin levels following a meal, it is possible to identify the presence of a metabolic disorder and determine if there is a defect in islet function, both of which are associated with diabetes. Insulin, a peptide hormone, is primarily produced by the beta cells of the pancreatic islets and serves as the body's main anabolic hormone. Its role involves regulating the metabolism of carbohydrates, fats, and proteins by facilitating the absorption of glucose from the bloodstream into the liver, adipose tissue (fat), and skeletal muscle cells.

BMI: Body mass index (BMI), a measure of obesity and health, is commonly used in statistical analysis. The degree of obesity cannot be judged directly by the absolute value of the weight, and it is naturally related to height. So, BMI is defined as the body mass divided by the square of the body height.
Diabetes Pedigree Function: The term used for this variable is DBF, which indicates the probability of developing diabetes depending on one's familial background (Joshi & Dhakal, 2021, 1-7).

Age: Age (years) the range in the dataset is from 21 to 81.

Outcome: Classification variable where 0 means that a female does not have Type II diabetes, and a 1 indicates the participant has Type II diabetes.

No.	Attributes	Attribute Type	Description

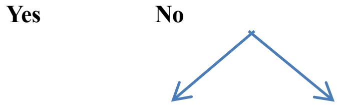
1	Pregnancies	Numerical	Number of times a Pima Indigenous female got pregnant
2	Glucose	Numerical	In an oral glucose tolerance test, plasma glucose concentration measured over 2 hours.
3	Blood Pressure	Numerical	Diastolic blood pressure (mm Hg)
4	Skin Thickness	Numerical	Thickness of Triceps skin fold (mm)
5	Insulin	Numerical	2-Hour serum insulin (mu U/ml)
6	BMI	Numerical	Body Mass Index (weight in kg/(height in m)²)
7	Diabetes Pedigree Function	Numerical	Diabetes pedigree function
8	Age	Numerical	Age in years
9	Outcome	Binary (0,1)	0 means that a female does not have Type II diabetes, and a 1 indicates the participant has Type II diabetes

Proposed Model: Logistic regression:

Within this dataset, the dependent variable "Outcome" exclusively consists of numerical values, specifically 0 and 1. As a result, logistic regression emerges as the most straightforward approach to utilize. Logistic regression serves the purpose of forecasting the likelihood of certain conditions transpiring in binary scenarios, such as yes/no or A/B situations. It enables the prediction of the probability of a categorical response transpiring based on the influence of one or more predictor variables.

Logistic regression surpasses discriminant analysis in its capability to analyze various categorical response variables due to its adaptability and versatility. Unlike discriminant analysis, which assumes the normality of all independent variables, logistic regression does not require this assumption. The fundamental concept of logistic regression revolves around a categorical dependent variable Y being regressed upon a set of p independent metric or binary variables $X1, X2, ..., Xp$. Examples of Y can include passing or failing an exam, being ill, or winning a prize. Logistic regression encompasses three types: binary logistic regression, multinomial logistic regression, and ordinal logistic regression. In our study, we will solely focus on binary logistic regression since the dependent variable "Outcome" in the dataset only has two possible values: "0" and "1" (Huang, 2021).

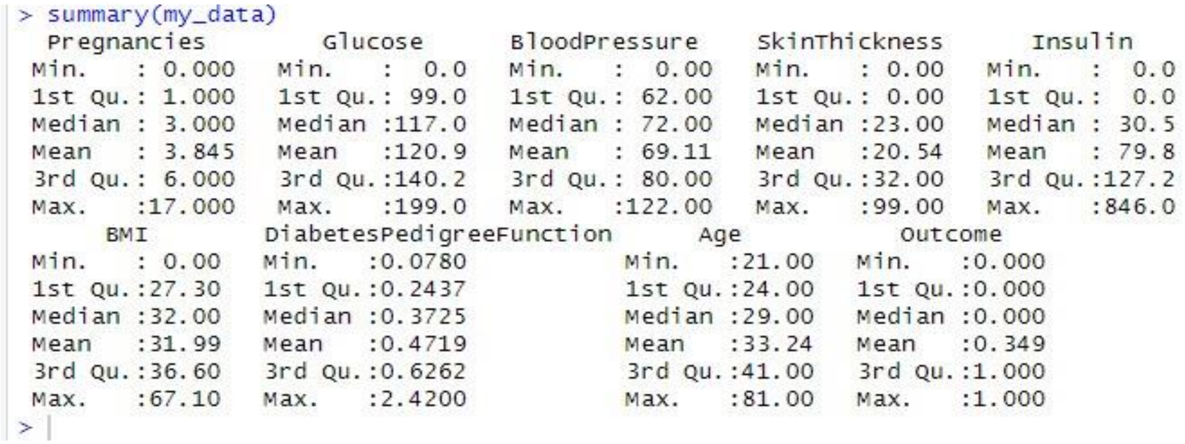
Figure 1: *Diabetes Test Result*



Analysis and Result

Step 1: Summarize data in table 2 and 3.

Table 2: Description of dataset summary which contains the following information about 768 individuals:



```
> summary(my_data)
  Pregnancies        Glucose       BloodPressure     SkinThickness      Insulin
 Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.   : 0.00   Min.   :  0.0
 1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.: 0.00   1st Qu.:  0.0
 Median : 3.000   Median :117.0   Median : 72.00   Median :23.00   Median : 30.5
 Mean   : 3.845   Mean   :120.9   Mean   : 69.11   Mean   :20.54   Mean   : 79.8
 3rd Qu.: 6.000   3rd Qu.:140.2   3rd Qu.: 80.00   3rd Qu.:32.00   3rd Qu.:127.2
 Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00   Max.   :846.0
      BMI        DiabetesPedigreeFunction      Age           Outcome
 Min.   : 0.00   Min.   :0.0780           Min.   :21.00   Min.   :0.000
 1st Qu.:27.30   1st Qu.:0.2437           1st Qu.:24.00   1st Qu.:0.000
 Median :32.00   Median :0.3725           Median :29.00   Median :0.000
 Mean   :31.99   Mean   :0.4719           Mean   :33.24   Mean   :0.349
 3rd Qu.:36.60   3rd Qu.:0.6262           3rd Qu.:41.00   3rd Qu.:1.000
 Max.   :67.10   Max.   :2.4200           Max.   :81.00   Max.   :1.000
> |
```

.

Table 3: Summary of variables

summary: summary function is applied to each variable	min: represents the minimum value
1st Qu: represents the first quarter	median: represents the median value
mean: represents the mean value	3rd Qu: represent the third quarter
max: represents the maximum value	

Check the normality distribution for the variables:

By the central limit theorem, we know that no matter what distribution things have, the sampling distribution tends to be normal if the sample is large enough (n > 30) Ref CLT.

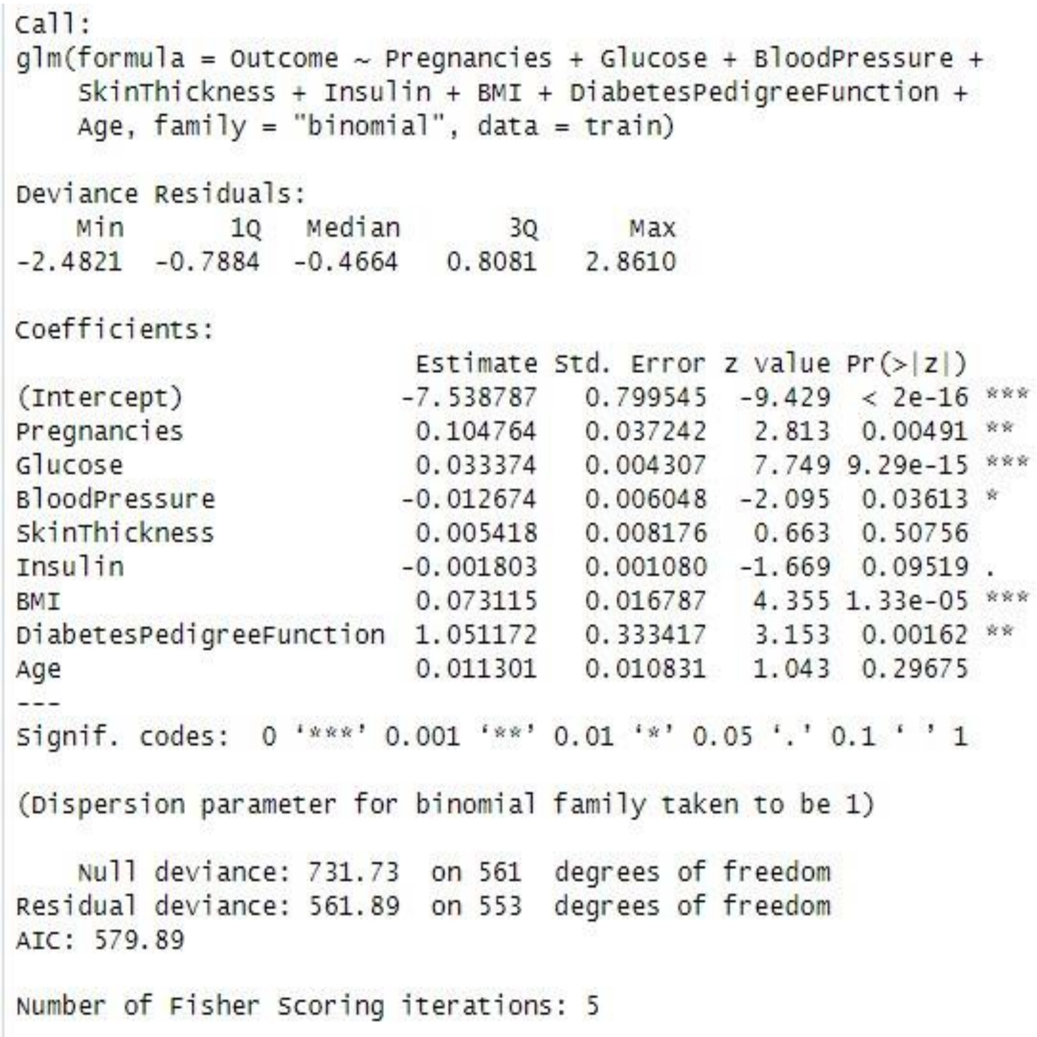
Step 2: Create Training and Test Samples

Next, to split the dataset into a training set to *train* the model on and a testing set to *test* the model. The data is split into 70% training data set and 30% testing data set.

Step 3: Fit the Logistic Regression Model

Next, to use the **glm** (general linear model) function and specify family =" binomial" so that R fits a logistic regression model to the dataset, table 4.

Table 4. The logistic regression model.



```
Call:
glm(formula = Outcome ~ Pregnancies + Glucose + BloodPressure +
    SkinThickness + Insulin + BMI + DiabetesPedigreeFunction +
    Age, family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4821  -0.7884  -0.4664   0.8081   2.8610

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -7.538787   0.799545  -9.429  < 2e-16 ***
Pregnancies               0.104764   0.037242   2.813  0.00491 **
Glucose                   0.033374   0.004307   7.749 9.29e-15 ***
BloodPressure            -0.012674   0.006048  -2.095  0.03613 *
SkinThickness             0.005418   0.008176   0.663  0.50756
Insulin                  -0.001803   0.001080  -1.669  0.09519 .
BMI                       0.073115   0.016787   4.355 1.33e-05 ***
DiabetesPedigreeFunction  1.051172   0.333417   3.153  0.00162 **
Age                       0.011301   0.010831   1.043  0.29675
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 731.73  on 561  degrees of freedom
Residual deviance: 561.89  on 553  degrees of freedom
AIC: 579.89

Number of Fisher Scoring iterations: 5
```

The coefficients presented in the output provide information about the average change in log odds of the "Outcome" variable. They represent the expected change in the log odds relative to a one-unit change in the corresponding variable. For instance, if we consider the variable "Pregnancies," a coefficient value of 0.104764 suggests that, on average, there is an increase of 0.104764 in the log odds of the "Outcome" variable for each additional unit of pregnancies, while controlling for other variables in the model. By exponentiation this coefficient (Exp(0.104764) ≈ 1.11), we find that each extra pregnancy is associated with an approximately 11% increase in the odds of having diabetes, assuming that all other variables in the model are held constant. This interpretation implies that the likelihood of having diabetes is expected to rise by approximately 11% for each additional pregnancy, taking into account the effects of other variables not included in the model.

The p-values in the output also give us an idea of how effective each predictor variable is at predicting the probability of diabetes (table 5):

Table 5: The P-value

Attribute	**P-value**
Pregnancies	**0.00491**
Glucose	**9.29e-15**
Blood Pressure	**0.03613**
Skin Thickness	**0.50756**
Insulin	: **0.09519**
BMI	**1.33e-05**
Diabetes Pedigree Function	**0.00162**
Age	**0.29675**

Based on the table provided, it appears that the variables "Pregnancies," "Glucose," "Blood Pressure," "Insulin," "BMI," and "Diabetes Pedigree Function" are considered important predictors. This conclusion is drawn from the observation that these variables have low p-values. A low p-value typically indicates that the relationship between the predictor variable and the outcome variable is statistically significant. Therefore, these variables are more likely to have a meaningful impact on predicting the outcome of interest, which in this case is diabetes.
On the other hand, the variables "Skin Thickness" and "Age" are noted as not being statistically significant in the model. This means that there is insufficient evidence to suggest a significant relationship between these variables and the outcome variable, at least based on the current model and dataset.

The equation of a logarithmic regression model takes the following form:

log (P (Y)/ (1-p (Y))) = XB

where:

Y: The response or the dependent variable
X: The predictor variables
B: The regression coefficients that describe the relationship between X and Y

Using the coefficients from the output table, its shown that the fitted logarithmic regression equation is:

log (P (Outcome)/ (1-p (Outcome))) = -7.5387+0.1047* (pregnancies)+0.0333* (glucose)-0.0126*(Blood Pressure) +0.0054*(Skin Thickness)-0.0018* (Insulin)+0.0731* (BMI)+1.0511*(Diabetes Pedigree Function) +0.0113*(Age)

Based on the values of the predictor variables, the following equation can be used to predict the value of the dependent variable (also referred to as the responder variable), denoted as "y": y = b0 + b1x1 + b2x2 + ... + bn*xn
In this equation, "b0" represents the intercept term, and "b1" through "bn" represent the coefficients associated with each predictor variable (x1 to xn). By substituting the specific values of the predictor variables into the equation, you can calculate the predicted value of the dependent variable, y.
It's important to note that the coefficients (b1 to bn) are estimated from the model fitting process and represent the average change in the dependent variable associated with a one-unit change in the corresponding predictor variable, while holding other variables constant.

Assessing Model Fit:

To compute a metric known as McFadden’s R2 as a way to assess how well a model fits the data, which ranges from 0 to just under 1. Values close to 0 indicate that the model has no predictive power. In practice, and as a rule of thumb is that McFadden’s R2 from 0.2 to 0.4 indicates very good model fit (McFadden, 1974). To compute McFadden’s R2 for the suggested model using the **pR2** function from the pscl package:

```
        pscl:       pR2(model)["McFadden"]
fitting null model for pseudo-r2
        McFadden            0.2321095
```

so the model fits the data very well.

Variable Importance:

To compute the importance of each predictor variable (table 6) in the model by using the varImp function from the caret package:

caret: varImp(model)

Table 6. Variable importance

Pregnancies	2.8130447	Glucose	7.7485730
BloodPressure	2.0954334	Skin Thickness	0.6626464
Insulin	1.6686422	BMI	4.3554252
Diabetes Pedigree Function	3.1527208	Age	1.0434181

Higher values indicate more importance. These results match up nicely with the p-values from the model. Glucose is the most important predictor variable, followed by BMI, Diabetes Pedigree Function, Pregnancies, Blood Pressure, Insulin.

VIF Values:

To calculate the VIF values (table 7) of each variable in the model to see if multicollinearity is a problem:

Variables	**VIF Values**
Pregnancies	1.440816
Glucose	1.269647
BloodPressure	1.192898
Skin Thickness	1.620979
Insulin	1.590187
BMI	1.252771
Diabetes Pedigree Function	1.027103
Age	1.541968

As a rule of thumb, VIF values above 5 indicate the existence of multicollinearity. Since none of the predictor variables in our models have a VIF over 5, It is assumed that multicollinearity is not an issue in the suggested model.

Step 4: Use the Model to Make Predictions using the Test Data

Once the logistic regression model is fitted, it is used to make predictions about whether or not an individual will have a diabetes based on the variables suggested in the model. The following code is used to calculate the probability of diabetes for every individual in our test dataset: #calculate probability of diabetes for each individual in test dataset predicted <- predict (model, test, type="response") **Step 5: Model Diagnostics:**

To evaluate the performance of the suggested logistic regression model on the test dataset, a common approach is to set a threshold for classification. By default, if the predicted probability of the outcome variable is greater than 0.5, the individual is predicted to have the outcome. However, it is worth noting that the choice of threshold can have an impact on the model's performance (**Long et al., 1993, 74-97).**

The predictions on the test datasets are made, create a confusion matrix with threshold value = 0.5 > confusion Matrix (test $Outcome, predicted)

```
        0    1
```

0	125	24
1	13	44

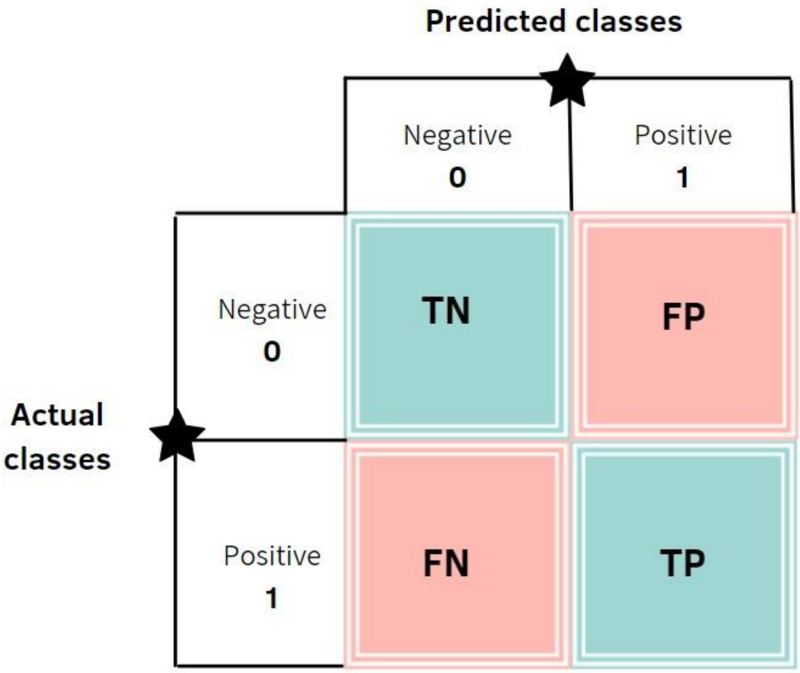


Figure.2 Confusion Matrix

True Positive (TP) is an outcome where the model correctly predicts the positive class.

True Negative (TN) is an outcome where the model correctly predicts the negative class.

False Positive (FP) is an outcome where the model incorrectly predicts the positive class.

False Negative (FN) is an outcome where the model incorrectly predicts the negative class.

Also to calculate the sensitivity (also known as the “true positive rate”) and specificity (also known as the “true negative rate”) along with the total misclassification error (which tells us the percentage of total incorrect classifications):

sensitivity	0.65
specificity	0.91
total misclassification error rate	0.18
Total accuracy	0.82

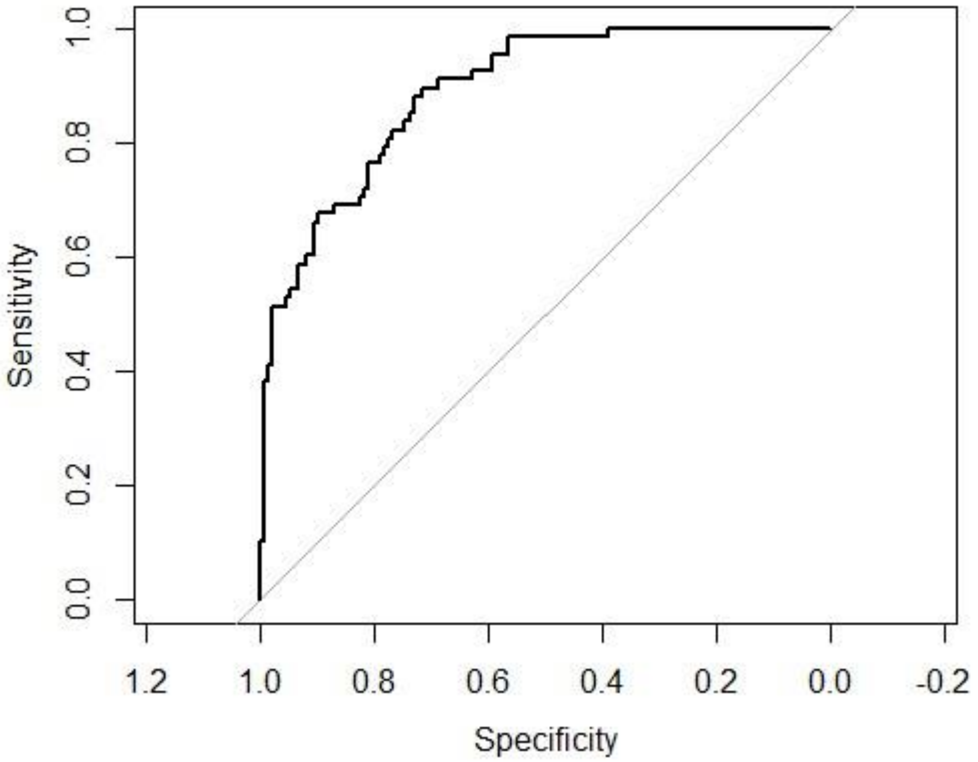


Figure.3 ROC curve

AUC = 0.890665

So it's shown that the AUC is **0.8906**, which is high. This indicates that our model does a good job of predicting whether or not an individual will be diabetic.

Conclusion and Recommendations

Detecting diabetes poses a significant obstacle in the healthcare industry. This research focused on predicting diabetes in women using the Pima Indigenous diabetes dataset specifically designed for female subjects. The dataset included various attributes such as pregnancies, glucose levels, blood pressure, skin thickness, insulin levels, BMI, diabetes pedigree function, and age, which were utilized for the prediction process. Logistic regression algorithms were employed for this purpose. The outcomes revealed that Logistic Regression achieved an accurate classification rate of 82.03% for the dataset.

Nevertheless, it is important to note that this study focused exclusively on the Pima Indigenous diabetes dataset for women, without incorporating additional diabetes datasets or comparing the accuracy with other machine learning techniques. Furthermore, the dataset used in this study had a relatively small size and a limited number of instances. Consequently, for future research, it is advisable to consider larger and more diverse datasets, as well as explore the application of various machine learning algorithms to enhance the accuracy and robustness of the diabetes prediction models.

Furthermore, considering the limited size of the dataset, which contained only 768 records, it would be prudent to gather more comprehensive data if the research were to be expanded in the future. It is also essential to include records of men in the dataset to ensure a more inclusive analysis. Additionally, incorporating additional features such as daily nutrition and exercise logs would provide valuable insights. By accessing a wider range of data, both men and women could enhance their ability to take necessary precautions and mitigate the risk of developing Type II diabetes.

With the advancement of technology, the future is expected to be characterized by increased computerization, networking, and reliance on computing equipment and technologies. In light of this, it would be advantageous for medical professionals to aim for the implementation of an Intelligent Medical Diagnostic System for Diabetes. Such a system would greatly enhance the accuracy and quality of diagnoses and medical care provided to patients with type 2 diabetes. By harnessing the power of intelligent algorithms and advanced computing capabilities, healthcare practitioners can significantly improve their diagnostic processes and overall patient outcomes.

References

[1]. Abu Duma, Haider. (2019). Using Logistic Regression and Discriminant Analysis Techniques for Factors Affecting Heart Diseases Infection. Ph. D. Dissertation, Sudan University for Science and Technology.

[2]. Alajlan, Abrar. (2021). A Model-Based Approach for an Early Diabetes Prediction Using Machine Learning Algorithms. Turkish Journal of Computer and Mathematics Education. 12 (3), 3957-3965.

[3]. Al-Saftawi, Yahya; Harz, Hussam; Rafi, Ahmed; Hijazi, Musbah. (2021). Predicting Diabetes Using JustNN. International Journal of Academic Health and Medical Research. 5 (3), 61-70.

[4]. American Diabetes Association. (2009). Diagnosis and classification of diabetes mellitus. Diabetes Care. 33 (Suppl. 1), 62–67.

[5]. Bhar, Haroun & Abu Sadah, Abdul Hadi. (2018). Uses of Logistic Regression to Determine Important Factors Affect Diabetes Meletus in Gaza Strip, Palestine. Journal of Al-Azhar University – Gaza, Palestine: Social Sciences Series. 20 (3), 231-261.

[6]. Brownlee, Jason. (2020). Logistic Regression for Machine Learning. Machine Learning Mastery. https://machinelearningmastery.com/logistic-regression-for-machine-learning/.

[7]. Cahyani, Qatrunnada; Finandi, Mochammad; Rianti, Jathu; Rianti, Arianti, Devi & Putra, Arya. (2022). Diabetes Risk Pre-diction Using Logistic Regression Algorithms. Journal of Machine Learning and Artificial Intelligence. 1 (2), 107-114.

[8]. Cho, N. H., Shaw, J. E., Karuranga, S., Huang, Y., da Rocha Fernandes, J. D., Ohlrogge, A. W. & Malanda, B. (2018). IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045. Diabetes research and clinical practice, 138, 271–281.

[9]. Daghistani, Tahani & Alshammari, Riyad. (2016). Diagnosis of Diabetes by Applying Data Mining Classification Techniques Comparison of Three Data Mining Algorithm. International Journal of Advanced Computer Science and Applications, 7 (7), 329-332.

[10]. Dhage, Sandhya, & Raina, Charanjeet, (2016). A review on Machine Learning Techniques. International Journal on Recent and Innovation Trends in Computing and Communication, 4 (3),

[11]. 395-399.

[12]. Fatima, Meherwar and Maruf, Pasha. (2017). Survey of machine learning algorithms for disease diagnosis. Journal of Intelligent Learning Systems and Applications. 9 (1), 1-16.

[13]. Huang, Ruodi. (2021). Prediction of Pima Indians Diabetes with Machine Learning Algorithms. MA Thesis, University of California, Los Angeles.

[14]. Joshi, Ram & Dhakal, Chandra (2021). Predicting Type 2 Diabetes Using Logistic Regression and Machine Learning Approaches. International Journal of Environmental Research and Public Health. 18 (14), 1-17.

[15]. Khalel, Intisar. 2016. Uses of Logistic Technique to Determine Factors Affect Marriage Delay in Saudi Arabia. Journal of Al-Sharjah University: Social and Human Sciences Series. 13 (2), 220- 246.

[16]. Lai, Hang; Huang, Huaxiong; Keshavjee, Karim; Guergachi, A; & Gao, Xin. (2019). Predictive models for diabetes mellitus using machine learning techniques. BMC Endocrine Disorders, 19(1), 1-9.

[17]. Long, W. J., Griffith, J. L., Selker, H. P., & D'agostino, R. B. (1993). A comparison of logistic regression to decision-tree induction in a medical domain. Computers and Biomedical Research, 26(1), 74-97.

[18]. Magoulas, George D & Prentza, Andriana. (2001). Machine learning in medical applications., in Lecture in Computer Science. Book Series. LUNAI, 2049. 300-307.

[19]. Murea, Mariana; Ma, Lijun & Freedman, Barry. (2012). Genetic and environmental factors associated with type 2 diabetes and diabetic vascular complications. Rev Diabet Stud. 2012 spring; 9(1), 6-22. https://my.clevelandclinic.org/health/diseases/16618-diabetes-insipidus.

[20]. Orabi, Karim; Kamal, Yasser; & Rabah, Thana. (2016). Early Predictive System for Diabetes. Industrial Conference on Data Mining: Application and Theoretical Aspects. 420-427. https://link.springer.com/chapter/10.1007/978-3-319-41561-1_31.

[21]. Rahimloo, Parastoo & Jafarian, Ahmad. (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.

[22]. Rajendra, Piryanka & Latifi, Shahram. (2021). Prediction of diabetes using logistic regression and ensemble techniques. Computer Methods and Programs in Biomedicine Update. 1, 1-8. https://www.sciencedirect.com/science/article/pii/S2666990021000318

[23]. World Health Organization. (2019). Classification of Diabetes Mellitus. file:///C:/Users/Majd/Dropbox/My%9789241515702-eng.pdf.