JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Prediction of Type-2 Diabetes using Logistic Regression and selection of optimal Normalization and Data Reduction Technique

1. Normalization and Data Reduction.

[1] Abhishek N, [2] Piyush Raj
[1]Student, [2]Student
[1]REVA Academy for Corporate Excellence,
[1]REVA University, Bengaluru, India

Abstract: This study aims to utilize machine learning, specifically logistic regression, to predict an individual's likelihood of having diabetes based on medical data, addressing a pressing global health concern. In addition to logistic regression, in this study we plan to explore techniques of scaling and varying range of train-test data ratios for each scaling technique and compare performance and accuracy of each scaler to identify best performing scaler and optimal train-test split ratio.

Index Terms **- Type 2 Diabetes, Logistic Regression, Min Max Scaling, Standardization, Robust Scaling, Divide by max.**

I. INTRODUCTION

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 (Rajendran & 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).

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).

Scaling: Scaling will help standardize the features in the given dataset, by adjusting the range of the data to fit within specific scale, which is essential and crucial for many machine learning algorithms.

Train-Test Split is a specific form of random sampling used to divide the data into training and testing sets, often with stratification to ensure balanced class distributions in both sets. Train-Test Split is one of the essential steps in the data preprocessing stage to ensure robust model training and evaluation.

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 m2).
7.	Diabetes Pedigree Function	Numerical	Diabetes pedigree function.
8.	Age	Numerical	Age in years.
9.	Outcome	Outcome	0 means that a female does not have Type II diabetes, and a 1 indicates the participant has Type II diabetes.

Figure 1: *Diabetes Test Result* - Yes No

Proposed Model: Logistic regression:

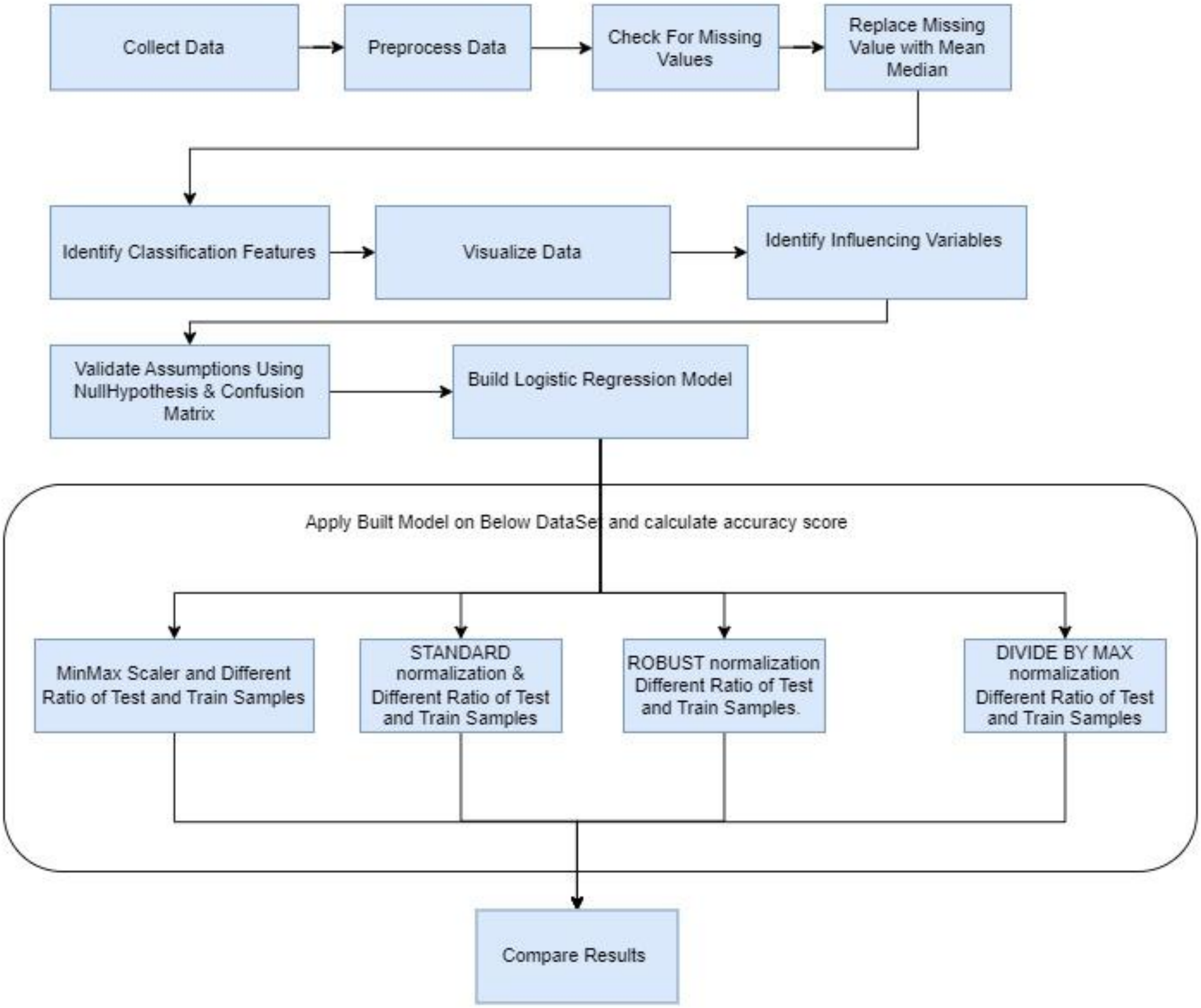


Figure 2: *Proposed Process flow*

The first step of Machine learning is exploring dataset, check data quality by verifying missing values, in the dataset used in this paper there are no missing values. So, there is no need to treat missing values.

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).

In this example we try to determine the feature in the dataset which has highest influence on the outcome by plotting density plot of each feature and validate our Hypothesis using Null Hypothesis.

Once Hypothesis is validated, we would like to explore the accuracy of Logistic Regression model on various scaling Normalization techniques and for each scaling technique, test the model on different ratio of Training samples and Test Samples.

Test Logistic Regression model on below use cases

Scaling Technique	Training/Test Dataset
Min Max Scaling	10% -Train, 90%-Test 20% -Train, 80%-Test 30% -Train, 70%-Test 40% -Train, 60%-Test
Standard Scaling	10% -Train, 90%-Test 20% -Train, 80%-Test 30% -Train, 70%-Test 40% -Train, 60%-Test
Robust Scaling	10% -Train, 90%-Test 20% -Train, 80%-Test 30% -Train, 70%-Test 40% -Train, 60%-Test
Divide by Max	10% -Train, 90%-Test 20% -Train, 80%-Test 30% -Train, 70%-Test 40% -Train, 60%-Test

Analysis and Result:

Step 1: Description of dataset summary which contains the following information and importing required libraries.

```
#import required libraries

#data mgmt libraries
import pandas as pd
import numpy as np

#data test libraries
from scipy import stats

#data viz libraries
import matplotlib.pyplot as plt
import seaborn as sns

#ignoring the warnings
import warnings
warnings.filterwarnings('ignore')

#importing user defined functions from functions.py
from functions import log_reg_minmax, log_reg_standard, log_reg_robust, log_reg_max
```

```
#Read the data
df = pd.read_csv('C:\\Repo\\AI08-M02\\Statistical-foundations\\data\\diabetes2.csv')
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
614	11	138	74	26	144	36.1	0.557	50	1
304	3	150	76	0	0	21.0	0.207	37	0
584	8	124	76	24	600	28.7	0.687	52	1
722	1	149	68	29	127	29.3	0.349	42	1
446	1	100	72	12	70	25.3	0.658	28	0
594	6	123	72	45	230	33.6	0.733	34	0
512	9	91	68	0	0	24.2	0.200	58	0
757	0	123	72	0	0	36.3	0.258	52	1
217	6	125	68	30	120	30.0	0.464	32	0
550	1	116	70	28	0	27.4	0.204	21	0

Step 2: Five-point summary of the data.

```
#five point summary of the data
df.describe()
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

Step 3: Check for missing values

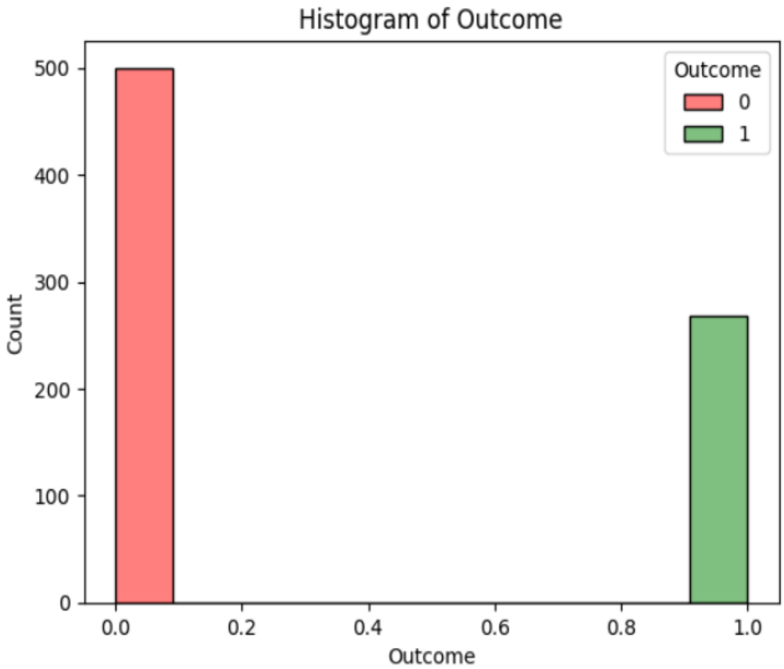
```
df_diabetics.isnull().sum() # check for null data in dataset
```

```
Pregnancies                 0
Glucose                     0
BloodPressure               0
SkinThickness               0
Insulin                     0
BMI                         0
DiabetesPedigreeFunction    0
Age                         0
Outcome                     0
dtype: int64
```

Step 4: Checking for Outcome frequency.

Outcome	Count
0	500
1	268

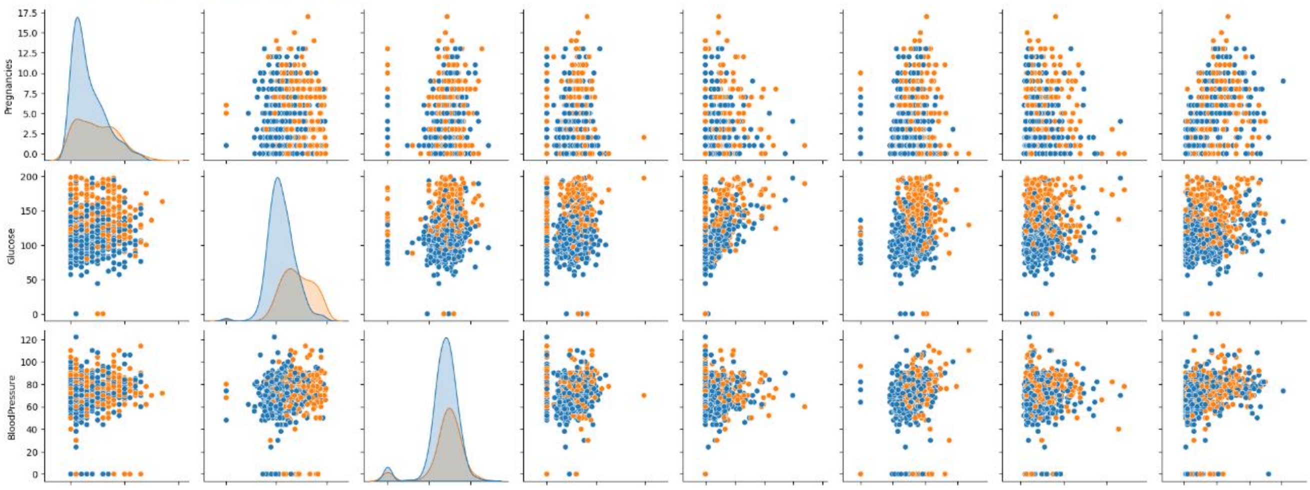
In the given dataset, 500 candidates are non-diabetic, and 268 candidates are diabetic.
Below graph shows visual plot of frequency distribution.



Step 5: Explore shape and relations - use the outcome column as colour.

```
#Explore shape and relations - use the outcome column as color
sns.pairplot(df, hue='Outcome')
```

```
<seaborn.axisgrid.PairGrid at 0x21ef0f86cc0>
```



Step 6:

Based on visual analysis, density plot of glucose and outcome, there is visual evidence and glucose is higher influencing factor in determining the outcome.

Outcome	Mean Glucose
0	109.98
1	141.257463

Independent T-Test performed using SciPy Tests

`T = -14.600060005973894, P= 8.935431645289913e-43`

Perform T-Test and Confusion matrix, confirm assumption, higher the value of Glucose, higher the probability of Candidate being Diabetic

Below table shows pivot distribution of, occurrence of diabetic and non-diabetic candidate with glucose level above 135.

Glucose_Level	High glucose	Low glucose
Outcome		
Diabetic	130	138
Non-Diabetic	57	443

Step 7: The accuracy score in the form of table for different scaling factors.

		Train	Test
Min Max Scaling	10% - Train, 90% - Test	0.802631579	0.718208092
	20% - Train, 80% - Test	0.738562092	0.738211382
	30% - Train, 70% - Test	0.77826087	0.737918216
	40% - Train, 60% - Test	0.755700326	0.793926247
Standardization	10% - Train, 90% - Test	0.736842105	0.764450867
	20% - Train, 80% - Test	0.751633987	0.775609756
	30% - Train, 70% - Test	0.769565217	0.773234201
	40% - Train, 60% - Test	0.785016287	0.752711497
Robust Scaling	10% - Train, 90% - Test	0.776315789	0.770231214
	20% - Train, 80% - Test	0.758169935	0.772357724
	30% - Train, 70% - Test	0.765217391	0.758364312
	40% - Train, 60% - Test	0.771986971	0.748373102
Divide by max	10% - Train, 90% - Test	0.723684211	0.690751445
	20% - Train, 80% - Test	0.784313725	0.700813008
	30% - Train, 70% - Test	0.791304348	0.750929368
	40% - Train, 60% - Test	0.745928339	0.770065076

Conclusion and Recommendations:

1. **Best Performing Scaler**:

   The **Min Max Scaler** with a train-test ratio of 0.1 and 0.9 yielded the highest accuracy score on train sample 80%. But accuracy dropped on test sample 71%

   The **Robust Scaler** with a train-test ratio of 0.1 and 0.9 yielded consistent accuracy of 77% for both test and train datasets.

   The **Divide by Max** with train-test ratio of 0.3 and 0.7 yielded accuracy score of 79% and 75% on test and train data set respectively

2. **Best Performing Train-Test ratio**

   The train-test ratio of **0.3 (30-70)** showed the best result across scalers, with average accuracy score over 77%. It indicates that this ratio provided sufficient data for training while maintaining a good generalization on the test set.

In conclusion, using the **RobustScaler** with a **10-90** train-test split ratio provides the best performance for logistic regression on this dataset. This combination yielded the highest mean accuracy Score, indicating a consistent performance of above accuracy for both test and train datasets.

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 Prediction 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.