International Journal of Scientific and Management Research
Volume 6 Issue 05 (May) 2023
ISSN: 2581-6888
Page: 34-48

Application of Logistic Regression Model in Prediction of Early Diabetes Across United States

I.Olufemi, C.Obunadike, A. Adefabi & D. Abimbola
[1,2,3,4] Department of Computer Science, Austin Peay State University, Clarksville, **USA**
DOI - http://doi.org/10.37502/IJSMR.2023.6502

Abstract

This study examines a case study and impact of predicting early diabetes in United States through the application of Logistic Regression Model. After comparing the predictive ability of machine learning algorithm (Binomial Logistic Model) to diabetes, the important features that causes diabetes were also studied. We predict the test data based on the important variables and compute the prediction accuracy using the Receiver Operating Characteristic (ROC) curve and Area Under Curve (AUC). From the correlation coefficient analysis, we can deduce that, out of the 16 PIE variables, only "Itching and Delayed healing" were statistically insignificant with the target variable (class) with a value of 83% and 33% respectively while "Alopecia and Gender/Sex" has a negative correlation with the target variable (class). In addition, the Lasso Regularization method was used to penalize our logistic regression model, and it was observed that the predictor variable "sudden_weight_loss" does not appear to be statistically significant in the model and the predictor variables "Polyuria and Polydipsa" contributed most to the prediction of Class "Positive" based on their parameter values and odd ratios. Since the confidence interval of our model falls between 93% and 99%, we are 95% confident that our AUC is accurate and thus, it indicates that our fitted model can predict diabetes status correctly.

Keywords: Machine Learning, Supervised Learning, Binomial Logistic Model, Early Diabetes Prediction.

1. Introduction

Diabetes is undoubtedly one of the most common diseases worldwide. It is one of the biggest health problems that affect millions of people across the world today. Many Machines Learning (ML) techniques have been utilized in predicting diabetes in the last couple of years and the ever-increasing complexity of this problem has inspired research scientist to explore other robust set of algorithms [2].

Logistic regression is a classification technique adopted by machine learning. It is also a statistical method applied when analyzing dataset with one or more PIE variables in other to determine the outcome. In other to find the best fit model that would describe the relationship between the DORT and PIE variables, logistic regression model is the best model that answers this puzzle [Bhuiyan].

This paper will help to solve the predominant challenges encountered in the health sector by applying machine learning and logistic regression model to accurately predict and identify diabetes at early stage. According to the Centers for Disease Control and Prevention [2], National Diabetes Statistics Report released in 2022. This report estimates that more than 130 million adults are living with diabetes or prediabetes in the United States [2]. The application of machine learning and artificial intelligence in the health sector will help to tackle this imminent problem aggressively and subsequently lead to drastic improvement in the health sector. The aims and objective of this journal is to help the United State health and medical sector in identifying diseases at a very early stage by using diabetes as a case study. In addition, based on the report from U.S Department of Health and Human Services, diabetes happens to be among the 15 leading causes of death in United States.

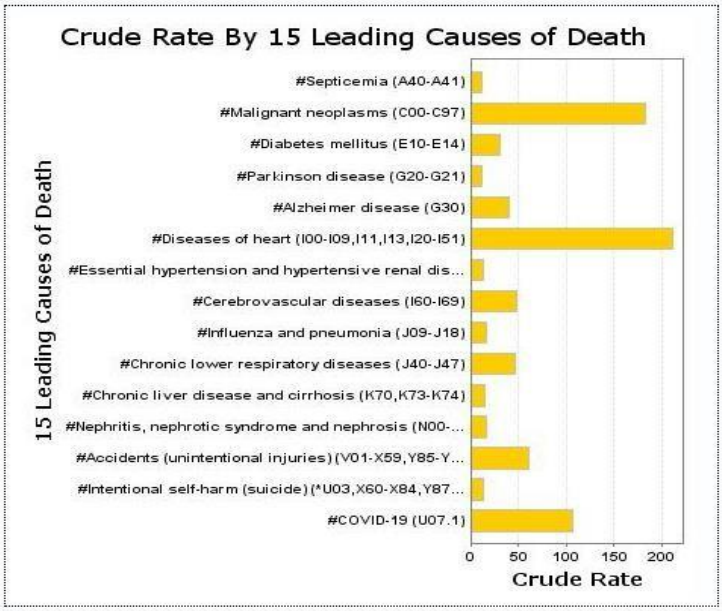

Fig. 1: 15 Leading Causes of Death in United States [2]

In the world today, diabetes is one of the frequent diseases that targets the elderly population. According to the International Diabetes Federation, 451 million people across the world were diabetic in 2019 [4]. The expectations are that this number will increase greatly to affect 693 million people in the coming 26 years. Diabetes is considered as a chronic disease associated with an abnormal state of the human body where the level of blood glucose is inconsistent due to some pancreas dysfunction that leads to the production of little or no insulin at all, causing diabetes of type 1 or cells to become resistant to insulin, causing diabetes of type 2.

The dataset used for this analysis was obtained from University of California, Irvine Machine Learning Repository. The dataset consists of multivariate data of 520 instances (rows or events) and 17 attributes (i.e., variables or columns). Out of the 17 variables, 'class' is assigned to be the target variable otherwise known as DORT or Y variable (i.e., dependent, observatory, response, or target variables) while the remaining 16 variables represents PIE or X variable (Predictor, Independent or explanatory variables).

The main cause of diabetes remains unknown, yet scientists believe that both genetic factors and environmental lifestyle play a major role in diabetes. Even though it's incurable, it can be

managed by treatment and medication. Individuals with diabetes face a risk of developing some secondary health issues such as heart diseases and nerve damage. Thus, early detection and treatment of diabetes can prevent complications and assist in reducing the risk of severe health problems.

2. Methodology

A systematic literature review was conducted to identify published studies on the early detection and prevention of diabetes [5]. We present in more detail about the use of data that is relevant with how to do early detection against Diabetes on an individual based on our survey activities and knowledge acquisition. The dataset was analyzed using R programming language and the first step adopted during the analysis was Exploratory Data Analysis (EDA). Exploratory data analysis is very vital because it allows us to have first insight about the association and relationship that exists between different variables. This could be often done by using some packages in R like plotly, naniar ggplot etc.

2.1 Descriptive Analysis of the Dataset

Out of the 17 variables, 'class' is assigned to be the target variable otherwise known as DORT or Y (dependent, observatory, response, or target variables) while the remaining 16 variables represents PIE or X (Predictor, Independent or explanatory variables). The target variable (class) was transformed from negative to 1 and positive to 0 (see Table 1).

Tables 1: Description and transformation of the data types (variables or features)

S/No	Variables		Initial Data Type (Idt)		Transformed data Type (Tdt)	
1	Age	20-65	Numeric/ Continuous		Numeric/ Continuous	
2	Sex/Gender	M/F	Categorical/Binary	Male/Female	Numeric/Binary	0/1
3	Polyuria	Yes/No	Categorical/Binary	Yes/No	Numeric/Binary	0/1
4	Polydipsia	Yes/No	Categorical/Binary	Yes/No	Numeric/Binary	0/1
5	Sudden Weight	Yes/No	Categorical/Binary	Yes/No	Numeric/Binary	0/1
6	weakness	Yes/No	Categorical/Binary	Yes/No	Numeric/Binary	0/1
7	Polyphagia	Yes/No	Categorical/Binary	Yes/No	Numeric/Binary	0/1
8	Genital thrush	Yes/No	Categorical/Binary	Yes/No	Numeric/Binary	0/1
9	Visual blurring	Yes/No	Categorical/Binary	Yes/No	Numeric/Binary	0/1
10	Itching	Yes/No	Categorical/Binary	Yes/No	Numeric/Binary	0/1
11	Irritability	Yes/No	Categorical/Binary	Yes/No	Numeric/Binary	0/1
12	Delayed healing	Yes/No	Categorical/Binary	Yes/No	Numeric/Binary	0/1
13	Partial paresis	Yes/No	Categorical/Binary	Yes/No	Numeric/Binary	0/1
14	Muscle stiffness	Yes/No	Categorical/Binary	Yes/No	Numeric/Binary	0/1

15	Alopecia	Yes/No	Categorical/Binary	Yes/No	Numeric/Binary	0/1
16	Obesity	Yes/No	Categorical/Binary	Yes/No	Numeric/Binary	0/1
17	Class	+/-	Categorical/Binary	Positive/Negative	Numeric/Binary	0/1

2.2 Checking for Missing Data

The anyNA() function was used to check for missing variables in our dataset. The outcome was "False". Thus, it implies that we did not have any missing data. In addition, we went further to visualize if there was any sort of missing data using the naniar package (see Fig.1). Based on figure 1 and 2, it clearly shows that we do not have any missing data.

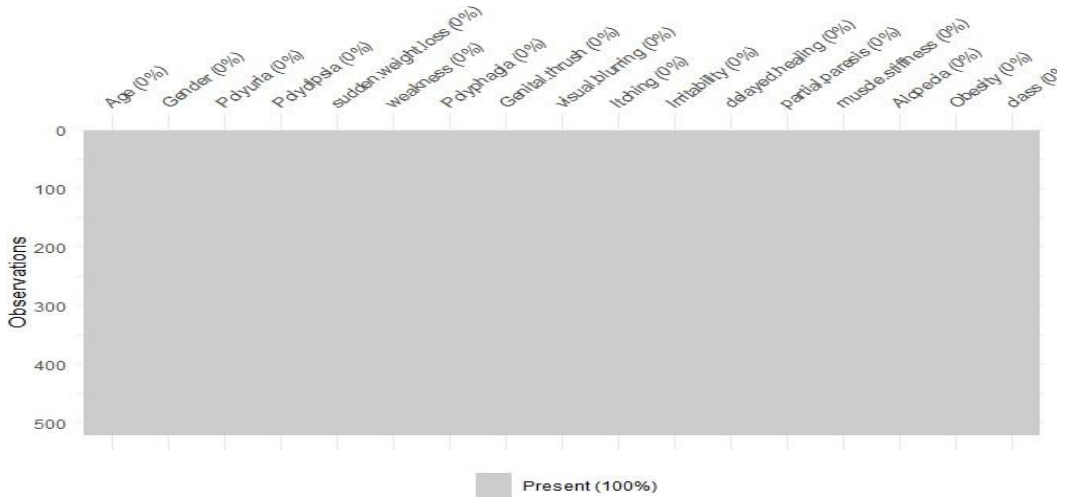

Figure 1: Barplot showing missing data using vis_mis() function in naniar package.

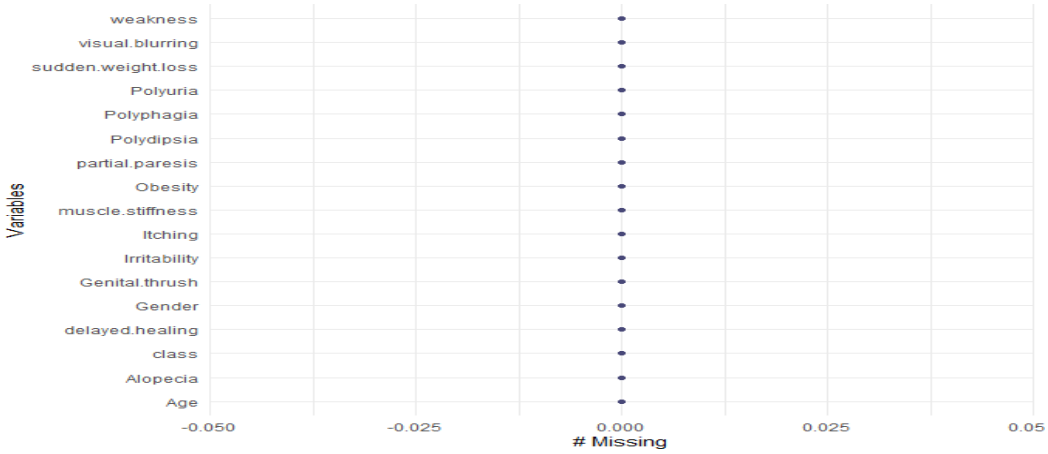

Figure 2: Graphical line plots showing missing using data using gg_miss_var() function in naniar package.

2.2.1 Intensive cross checking of other missing values

To ensure that our analysis and model would be free from errors. It is very important to thoroughly loop through the whole dataset to check for other missing values that may occur in other forms asides from "NA". From table 2 below, we could see that the ncom, nmiss and miss.prop shows that we do not have any missing values. This implies that our dataset is ready for use.

Tables 2: Iteration through the dataset using for loop to check for other missing values

Col.num	V.name	Mode	N.level	ncom	nmiss	Miss.prop
1	Age	numeric	51	520	0	0
2	Sex/Gender	numeric	1	520	0	0
3	Polyuria	numeric	1	520	0	0
4	Polydipsia	numeric	2	520	0	0
5	Sudden Weight	numeric	2	520	0	0
6	weakness	numeric	2	520	0	0
7	Polyphagia	numeric	2	520	0	0
8	Genital thrush	numeric	2	520	0	0
9	Visual blurring	numeric	2	520	0	0
10	Itching	numeric	2	520	0	0
11	Irritability	numeric	2	520	0	0
12	Delayed healing	numeric	2	520	0	0
13	Partial paresis	numeric	2	520	0	0
14	Muscle stiffness	numeric	2	520	0	0
15	Alopecia	numeric	2	520	0	0
16	Obesity	numeric	2	520	0	0
17	Class	numeric	2	520	0	0

2.2.2 Frequency distribution of target variable and box plots between target and predictor variable

Fig. 3 shows the frequency distribution of the target variable (class). It shows that the frequency distribution of the target variable “class”. Out of the 520 events or rows, 0.62 or 62% of the class are “Positive” while 0.38 or 38% are negative (see fig 3). Based on the frequency distribution of our target variables, it indicates that we have unbalanced classification. Thus, we would further investigate the association of the class variable (DORT) with other 16 variables (PIE). Since the target variable (class) is categorical and “Age” variable is continuous(discreet), Wilcoxon rank sum test was applied to check for their association. Based on the result, the p-value of 0.0124 indicates that there is an association between the two variables because the p-value is below the benchmark of 0.05. Furthermore, a correlation coefficient of -0.11 was seen between the target variable (class) and predictor (gender/sex) variable (see fig. 5). Table 3 shows the p-values and test-statistics of all other 16 predictor variables with the target (class) variable. From the result, it is obvious that both Itching and Delayed healing are unimportant variables because their p-values were higher than the benchmark. It is important to note that, asides from Age variable, chi-square was used to check for association between the remaining variables because they are categorical variables.

2.2.3 Correlation coefficients and association between variables

Correlation is a very strong statistical measurement. It helps to identify the association between two variables, and it ranges from -1 to +1. Based on our analysis, it could be seen that all the

16 predictor variables have an association with the *DORT* variable (class) except for *Itching* and *Delayed healing* (see fig. 5). In addition, Alopecia and Gender/sex have negative

$$r = \frac{\sum(x_i-\bar{x})\ (y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2\ \sum(y_i-\bar{y})^2}} \quad \text{..... Eqn. 1}$$

correlation with the target variable among the 16 predictor variables (see fig. 5). The association of the target variable on the predictor variables was observed (see Table 3). The marginal (bi-variate) associations between the class variable (DORT) and each predictor variables were done to classify the predictor variables into important and unimportant predictor variables (see Table 3 and Fig. 5).

r: correlation coefficient.

x_i: values of *PIE* (predictor, independent, or explanatory) variables in samples.

y_i: values of *DORT* (dependent, observatory, response target) variables in samples.

$\bar{x}$: mean values of *PIE* (predictor, independent, or explanatory) variables in samples.

$\bar{y}$: mean values of *DORT* (dependent, observatory, response target) variables in samples.

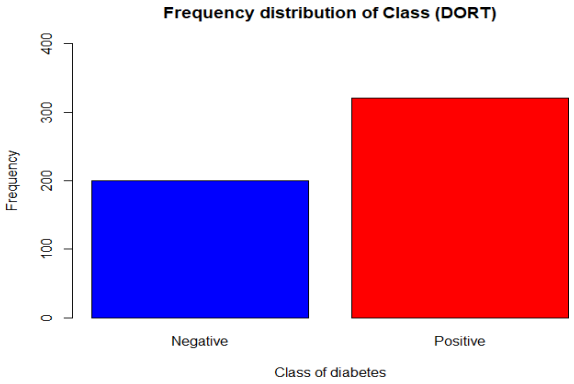

Fig 3: Frequency distribution of DORT (class)

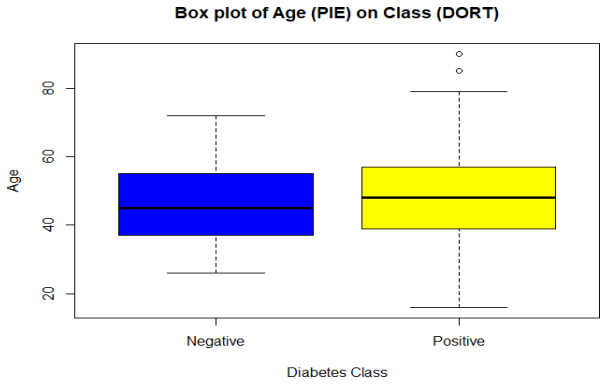

Fig 4: Box plot of Age (PIE) on Class (DORT)

Tables 3: Test-statistics and p-values between the variables

	Test-Statistic	P-value	Decision	H0	HA
Age	27834	0.0124	important		☐
Sex/Gender	103.03	0	important		☐
Polyuria	227.86	0	important		☐
Polydipsia	216.17	0	important		☐
Sudden Weight	97.29	0	important		☐
weakness	29.76	0	important		☐
Polyphagia	59.59	0	important		☐
Genital thrush	5.79	0.0161	important		☐
Visual blurring	31.80	0	important		☐

Itching	0.04	0.82975	unimportant	□	
Irritability	45.20	0	important		□
Delayed healing	0.96	0.32666	unimportant	□	
Partial paresis	95.38	0	important		□
Muscle stiffness	7.28	0.00694	important		□
Alopecia	36.06	0	important		□
Obesity	2.32	0.12711	important	□	

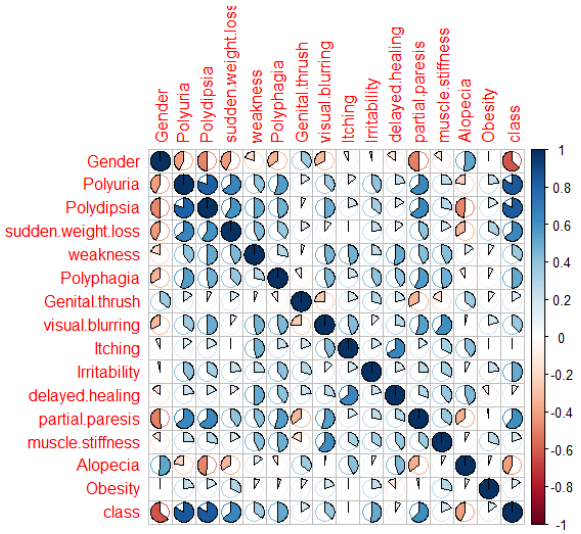

Fig 5: Correlation between DORT (target) and PIE variables

3. Results and Discussion

Logistic regression technique was applied to the train data to build a predictive model. Firstly, we adopted the lasso regularization (L1) with penalty to obtain the tuning parameter (λ) with cross validation. It is important to note that the L1 penalty is used for both variable selection and shrinkage, since it has the ability of forcing some of the coefficient estimates to be zero. Tables 5 represents important features (PIE variables) using the best predictive model with L1 penalty. Based on Table 5, it is obvious that out of the 16 predictor variables only "sudden_weight_loss" is not important predictors. The test data was predicted using the predictive model as well as the computation of its accuracy.

3.1 Dataset Partitioning

Since we have performed EDA and determined the important features (variables), we are ready to fit the model on our training data Therefore, we must split the dataset into training (70%) and testing data (30%). The logistic regression model that was used is the binomial due to the features of our dataset and target variable. In addition, for consistency of the partitioned dataset, we applied the set.seed(123) function for consistent values. The set.seed() function could take any values within the bracket.

3.2 Model Fitting

Since our target (DORT) variable is binary (Categorical), we are resorting to a classifier machine model (CMM). We will be using Logistic Regression model and Lasso as the

Regularization approach to predict Diabetes status of a patient. We chose Lasso Regularization approach because our focus is having a Parsimonious model that adequately explains the target (DORT) variable. We fit the Lasso

Logistic regression model with 200 sequences of turning parameter with lower limit 0.001 and upper limit 0.5 (see fig 6).

```
Lambda <- seq(0.0001, 0.5, length.out = 200)
L <- length(Lambda)
OUT <- matrix (0, L, 3)
library(glmnet)
for (i in 1:L){
  fit <- glmnet(x=D1.x, y=D1.y, family ="binomial", alpha =1,
                lambda=Lambda[i], standardize=FALSE, thresh = 1e-07, maxit=1000)
  pred <- predict(fit, newx=D1.x, s=Lambda[i], type="response")
  miss.rate <- mean(D1.y != (pred >=0.5))
  mse <- mean((D1.y - pred)^2)
  OUT[i, ] <- c(Lambda[i], miss.rate, mse)
}

head(OUT)
```

Fig 6: Sample code used for fitting the model

Table 4: Lambda's, misclassification rate and mean square error matrix

	[,1]	[,2]	[,3]
[1,]	0.000100000	0.1182796	0.08601410
[2,]	0.002612060	0.1102151	0.08859546
[3,]	0.005124121	0.1102151	0.09098334
[4,]	0.007636181	0.1129032	0.09312719
[5,]	0.010148241	0.1155914	0.09532075
[6,]	0.012660302	0.1155914	0.09790142

Although the best lambda to regularize the model is evaluated using the Means Square Error and Miss classification rate. The result of the first six rows of the Lambda's, miss classification rate and mean square error is printed above (see Table 4). Using MSE (Mean Square Error) as the evaluation metric. Our Best Lambda (tuning parameter) is 0.0004 (see Fig 7).

```
> lambda.best <- OUT[which.min(OUT[, 3]), 1]
> lambda.best
[1] 1e-04
```

Fig 7: Best tuning parameter based on Table 4

3.3 Lambda Vs Mean Square Error Plots

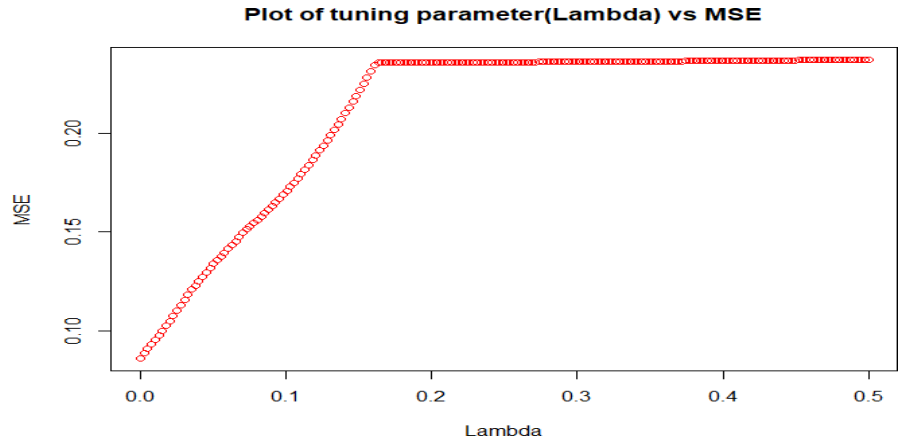

Fig 8: Showing plot of Lambda Vs MSE

Figure 8 shows as the tuning parameter (lambda) increases alongside with the MSE. Thus, it implies that the value of Lambda must kept minimal to have a low MSE value. In addition, at 0.18 lambda, the MSE values becomes uniform.

3.4 Final Model Fitting

Having gotten the best lambda. We fit the final Lasso Logistic regression model with the Training and Validation data pooled together (see fig. 9).

```
> fit.best <- glmnet (x=D1.x, y=D1.y, family ="binomial", alpha=1,    #LASSO
+                     lambda = lambda.best, standardize = FALSE,
+                     thresh = 1e-07, maxit=1000)
> names(fit.best)
 [1] "a0"          "beta"        "df"          "dim"         "lambda"      "dev.ratio"   "nulldev"
 [8] "npasses"     "jerr"        "offset"      "classnames"  "call"        "nobs"
> |
```

Fig 9: Showing sample code for final model fitting

Table 5: Coefficients of important predictors using LGR (l1) model

Variables	Coefficients
Age	0.06534024
Sex/Gender	-4.99548379
Polyuria	5.66066441
Polydipsia	6.02179976
Sudden Weight Loss	-
weakness	1.39414925
Polyphagia	1.38682283
Genital thrush	1.33510590
Visual blurring	0.13044201
Itching	-3.67583671
Irritability	1.70405358
Delayed healing	-0.45646626
Partial paresis	1.62380210
Muscle stiffness	-0.99207617
Alopecia	1.49216635
Obesity	-0.22411546

After fitting the model with the best lambda, only the variable "sudden_weight_loss" appears not to be important in the model (see Table 5).

3.5 Odds ratio of the predictor variables

According to the Centres for Disease Control and Prevention [2], odds ratio is the "measure of association" for a case-control study. It quantifies the relationship between an exposure and a disease in a case-control study. The odds ratio is calculated using the number of case-patients who did or did not have exposure to a factor and the number of controls who did or did not have the exposure. The odds ratio tells us how much higher the odds of exposure are among case-patients than among controls.

Generally, the intensity of the odds ratio is called the "strength of the association." The further away an odds ratio is from 1.0, the more likely it is that the relationship between the exposure and the disease is causal. For instance, an odds ratio of 1.25 is above 1.0, but is not a strong association while that of > 9.5 suggests a stronger association.

Table 6: Odds Ratio Indication and Implication [1]

Odds ratio	Indication	Implication
o 1.0 (or close to 1.0)	The odds of exposure among case-patients are the same as, or similar to, the odds of exposure among controls.	The exposure is not associated with the disease
o > 1.0	the odds of exposure among case patients are greater than the odds of exposure among controls	The exposure might be a risk factor for the disease.
o < 1.0	the odds of exposure among case patients are lower than the odds of exposure among controls.	The exposure might be a protective factor against the disease

The odds ratio of the important predictor variables was calculated using the exponential of the fit. Best and the beta (see Table 6 and 7).

Table 7: Odds Ratio of the Predictor Variables

	Odds Ratio	**Implications**	**Target Variable (class)** ***DORT***		**Likely to Occur**
			Positive (0)	Negative (1)	Y (0) / N (1) F (1) / M (0)
Age	1.067522	Age might be a risk factor for diabetes	□		Yes
Sex/Gender	0.006768446	Sex might be a protective factor for diabetes		□	Female
Polyuria	287.3395	Polyuria might be a risk factor for diabetes	□		Yes
Polydipsia	412.32	Polydipsia might be a risk factor for diabetes	□		Yes
weakness	4.031543	weakness might be a risk factor for diabetes	□		Yes
Polyphagia	4.002114	Polydipsia might be a risk factor for diabetes	□		Yes
Genital thrush	3.800398	Genital thrush might be a risk factor for diabetes	□		Yes
Visual blurring	1.139332	Visual blurring might be a risk factor for diabetes	□		Yes
Itching	0.0253282	Itching might be a protective factor for diabetes		□	No
Irritability	5.496182	Irritability might be a risk factor for diabetes	□		Yes
Delayed healing	0.6335184	Delayed healing might be a protective factor for diabetes		□	No
Partial paresis	5.072339	Partial paresis might be a risk factor for diabetes	□		Yes
Muscle stiffness	0.370806	Muscle stiffness might be a protective factor for diabetes		□	No
Alopecia	4.446718	Alopecia might be a risk factor for diabetes	□		Yes
Obesity	0.7992229	Obesity might not be associated with diabetes		□	Maybe/No

3.6 Model Evaluation

It is important to recall that model evaluation decides whether the model performs better. Therefore, it is critical to consider the model outcomes according to every possible evaluation method. Applying different methods can provide different perspectives. The following line of code was used to generate our AUC (see fig.10).

```
pred <- predict(fit.best, newx = D2.x, s =lambda.best, type="response")
library(cvAUC)
yobs <- D2.y
pred1 <- ifelse(pred>=0.5, 1, 0)
AUC <- ci.cvAUC(predictions = as.numeric(pred), labels =yobs, folds=1:NROW(D2.y))
auc.ci <- round(AUC$ci, digits = 3)
AUC
```

Fig. 10: Showing line of codes used to obtain the AUC (Area Under Curve)

Table 8: Results of the Model Evaluation

cvAUC	se	ci	confidence
0.9594332	0.0145292	0.9310 0.9879	0.95

From Table 8, the AUC of 0.9594 indicate that our fitted model has 95.9% ability to correctly classify diabetes status class positive or negative. The confidence interval also indicates the true AUC falls within the interval (0.9310, 0.9879). Therefore, we are 95% confident that our AUC is accurate.

3.7 Receiver Operating Characteristic (ROC) curve

The ROC curve below shows the trade-off between sensitivity (or TPR) and False Positive Rate (1 – Specificity). It further indicates that the model performs better against the benchmark (50%) with total area of 0.9594(95.9%).

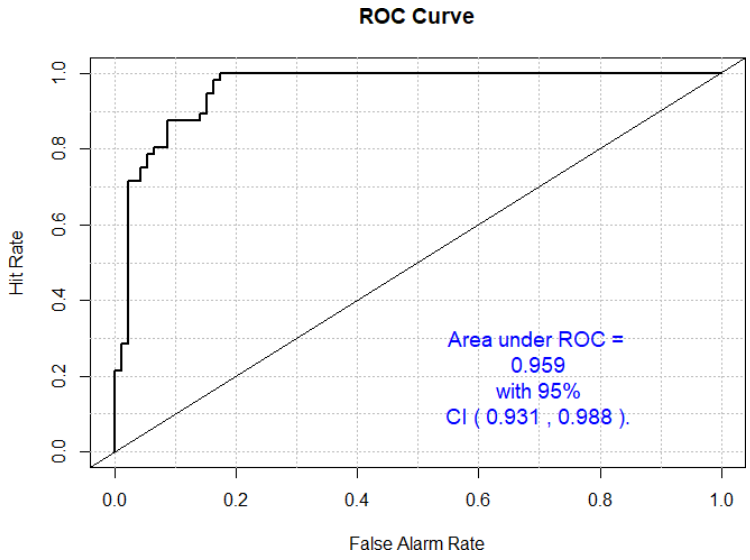

Fig. 11: Showing the Receiver Operators Curve of the Hit Rate Vs False Alarm

3.8 Confusion Matrix and statistics

From the result of the Confusion matrix, the target variable (class) positive is represented with the value "0" and the negative with value "1" (see Table 7). The Sensitivity (TPR) also Known as "Recall" has a value of 0.8696 (87%) thus it indicates that our model has a higher percentage of detecting Positive Diabetes class while it has a Specificity (TNR) value of 0.8750 (87.50%) further indicating that our model has a higher percentage of detecting negative diabetes class (see Table 9). In addition, the precision value of 92% indicates that our Model has a low false positive rate (i.e., less classification error or high ability to predict correctly positive or negative diabetes class). Furthermore, the F1 score of 0.8939 indicates that our model performs better. The higher the F1 score the better the performance of a binary classification (see Table 9).

Table 9: Confusion Matrix and Statistics

Confusion Matrix and Statistics	
Accuracy	0.8716
95%CI	(0.8068, 0.9209)
No Information Rate	0.6216)
P-Value [Acc > NIR]	1.225e-11
Kappa	0.7318
Mcnemar's Test P-Value	0.3588
Sensitivity	0.8696
Specificity (True Negative Rate/TNR)	0.8750
Pos Pred Value	0.9195
Neg Pred value	0.8033
Precision	0.9195
Recall (True Positive Rate/TPR)	0.8696
F1	0.8939
Prevalence	0.6216
Detection Rate	0.5405
Detection Prevalence	0.5878
Balanced Accuracy	0.8723
Positive Class	0
Precision	0.9195402

4. Conclusion

The analysis started with the Description of the dataset in terms of the sample size, data type of the predictor variables. We moved on to data cleaning of the dataset i.e., checking for missing values, outliers, and wrong records. We proceeded into performing exploratory data analysis of the dataset which involve visualizing the distribution of the target variable classification. Association of the predictor variables with the target variable using Correlation, Wilcoxon and chi-square was carried out to examine the strength of relationship between the predictor variables and the target. The next phase of the analysis was the model building. It started with dividing the dataset into training and test dataset. The dataset was trained with lasso penalized

logistics regression model where the lambda penalized parameter was first determined. After that, the best lambda was used to fit a Logistic regression model on the training dataset. In terms of the model performance metrics, we examined the model AUC (ROC curve), Recall, Precision and F1.

Lasso Regularization method was used to penalize our Logistic regression Model, it was observed that the predictor variable sudden_weight_loss does not appear to be statistically significant in the model. Predictor variables Polyuria and Polydipsa contributed most to the prediction of Class "Positive" based on their parameter values and odd ratios. The Gender predictor variable indicates male are more likely to have diabetes than female. It was also determined that our fitted model has an AUC of 95.9% with a Recall of 87%, precision of 92% and an F1 score of 89.4%. All this indicate our fitted model has a higher performance in explaining the target variable ("Diabetes class status"). Finally, based on the research questions related to diabetes symptoms we can determine the diabetes status of the respondent.

References

1) American Diabetes Association. Diagnosis and classification of diabetes mellitus. Diabetes Care 2009, 32, S62–S67. [CrossRef] [PubMed]
2) CDC 2022 National Diabetes Statistics Report: 2022 National Diabetes Statistics Report.
3) Center for Disease Control and Prevention. National Diabetes Fact Sheet: National Estimates and General Information on Diabetes and Prediabetes in the United States, 2011; US Department of Health and Human Services, Centers for Disease Control and Prevention: Atlanta, GA, USA, 2011; Volume 201, pp. 2568–2569.
4) Centre For Disease Control and Prevention: CDC Wonder https://wonder.cdc.gov/controller/datarequest/D76;jsessionid=F1696B2C464E3B34D922962F 0D4E
5) Hastie T, Tibshirani R, Friedman J (2008) The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2 Eds, Springer
6) IDF Diabetes Atlas 2022 Report. Available from https://diabetesatlas.org/
7) IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045
8) International Diabetes Federation: 02 November 2021 Affecting one in 10 adults https://www.idf.org/news/240:diabetes-now-affects-one-in-10-adults-worldwide.html
9) James G, Witten D, Hastie T, et al. (2013) An Introduction to Statistical Learning with Applications in R. Springer.
10) Kaur, H.; Kumari, V. Predictive modelling and analytics for diabetes using a machine learning approach. Appl. Comput. Inform. 2020, 18, 90–100. [CrossRef]
11) Machine Learning Repository. Early-Stage Diabetes Risk Data Set. Available from: https://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset
12) Maniruzzaman, M.; Rahman, M.; Ahammed, B.; Abedin, M. Classification and prediction of diabetes disease using machine learning paradigm. Health Inf. Sci. Syst. 2020, 8, 1–14. [CrossRef] [PubMed]

13) Salmonella in the Caribbean - 2013 Interpreting Results of Case-Control Studies https://www.cdc.gov/training/SIC_CaseStudy/Interpreting_Odds_ptversion.pdf
14) Saeedi, P.; Petersohn, I.; Salpea, P.; Malanda, B.; Karuranga, S.; Unwin, N.; Colagiuri, S.; Guariguata, L.; Motala, A.A.; Ogurtsova, K.; et al. Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas. Diabetes Res. Clin. Pract. 2019, 157, 107843. [CrossRef] [PubMed]
15) Sisodia, D.; Sisodia, D.S. Prediction of diabetes using classification algorithms. Procedia Comput. Sci. 2018, 132, 1578–1585. [CrossRef]