Binary Logistic Regression Analysis for Diabetes Diagnosis based on Blood Results: A Statistical Report

*Part B: Binary Logistic Regression Analysis

Saheli Dutta
x21246513student@ncirl.ie
National College of Ireland
Masters in Data Analytics

***Abstract*—This report presents a binary logistic regression analysis conducted on a dataset of blood sample details from diabetic patients collected at an Iraqi University Hospital in 2020. The objective is to develop a predictive model for diabetes diagnosis based on blood results. Descriptive statistics and visualizations are utilized to gain insights into the dataset, while the binary logistic regression model is built and evaluated. The report discusses the model-building steps, including data preprocessing and variable selection. Assumptions of the model are verified, and model performance and fit are assessed. The results highlight the potential for accurate diabetes diagnosis based on blood test variables. In the context of this study, binary logistic regression is employed to estimate the probability of a patient being diagnosed with diabetes or not, leveraging blood test variables.**

***Index Terms*—Diabetes diagnosis, Blood Samples Variables, Binary Logistic Regression, Descriptive Statistics, Predictive Modeling, Model performance evaluations, and Statistical Study.**

I. Introduction

Diabetes is a chronic metabolic disorder characterized by high blood glucose levels. Timely and accurate diagnosis is crucial for effective management and prevention of complications. Predictive models based on blood test variables have emerged as valuable tools for diabetes diagnosis.

This study utilizes a dataset that consists of detailed blood sample information from diabetic patients. The dataset includes parameters such as urea, creatinine ratio, average blood glucose levels, cholesterol, triglycerides, HDL, LDL, VLDL, BMI, and patient demographics.

The motivation behind this research is to enhance diagnostic accuracy and efficiency in diabetes diagnosis by developing a reliable predictive model using binary logistic regression. By analyzing the relationship between the blood test variables and diabetes status, significant predictors are identified, leading to the establishment of a robust diagnostic model.

The subsequent sections of this report delve into data exploration, preprocessing steps, model-building processes, evaluation of model performance, and the final results. The report concludes with a discussion of the findings, their implications, and future research prospects.

The study highlights the potential of blood test variables as reliable indicators for diabetes diagnosis, leading to accurate and timely diagnoses and improved treatment strategies. Additionally, the developed model contributes to existing diabetes diagnosis research and encourages further investigations.

II. Related Work

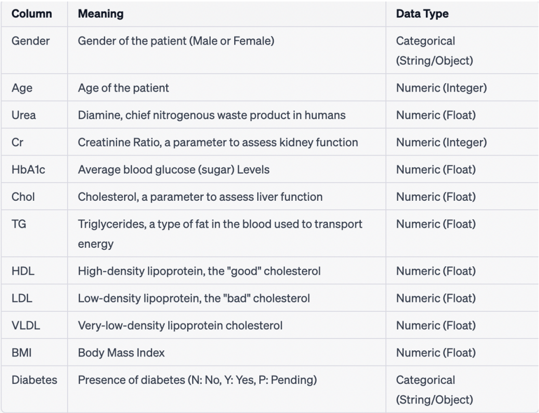
The study titled ”Prediction of Diabetes using Logistic Regression and Ensemble Techniques” by Priyanka Rajendra and Shahram Latifi focuses on the application of logistic regression, a commonly used classification model in machine learning, for predicting diabetes in patients. The authors explore the importance of early detection for effective diabetes management and prevention of complications. The study utilizes diagnostic measurements from two datasets: the PIMA Indians Diabetes dataset from the National Institute of Diabetes and Digestive and Kidney Diseases, and a dataset based on a study of rural African Americans in Virginia. The authors employ feature selection techniques to enhance the model's performance and utilize ensemble methods to improve prediction accuracy. The results show the highest achieved accuracy of approximately 78% for Dataset 1 and 93% for Dataset 2. This study underscores the effectiveness of logistic regression and emphasizes the significance of data preprocessing, feature selection, and the utilization of ensemble techniques in advancing prediction models for diabetes [1].

These insights are valuable for my own research, as they provide a foundation for developing an accurate prediction model for diabetes using similar techniques. By incorporating pre-processing of data and feature selection into the model design, it is possible to enhance the accuracy and performance of the predictive system. This study serves as a vital reference, emphasizing the importance of data preprocessing and highlighting the potential benefits of utilizing logistic regression in advancing my own research on diabetes prediction and management.

III. Methodology

A. Data Sets and Characteristics

The dataset used in the study consists of 1000 entries with 12 columns. The dataset contains the following columns (**Fig 1**) The dataset includes a mix of numerical and categorical variables.
The independent variables (features) include Gender, AGE,

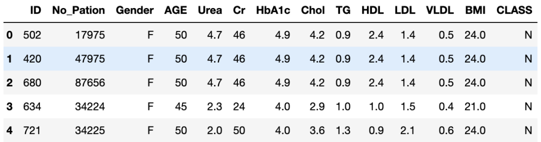


Column	Meaning	Data Type
Gender	Gender of the patient (Male or Female)	Categorical (String/Object)
Age	Age of the patient	Numeric (Integer)
Urea	Diamine, chief nitrogenous waste product in humans	Numeric (Float)
Cr	Creatinine Ratio, a parameter to assess kidney function	Numeric (Integer)
HbA1c	Average blood glucose (sugar) Levels	Numeric (Float)
Chol	Cholesterol, a parameter to assess liver function	Numeric (Float)
TG	Triglycerides, a type of fat in the blood used to transport energy	Numeric (Float)
HDL	High-density lipoprotein, the "good" cholesterol	Numeric (Float)
LDL	Low-density lipoprotein, the "bad" cholesterol	Numeric (Float)
VLDL	Very-low-density lipoprotein cholesterol	Numeric (Float)
BMI	Body Mass Index	Numeric (Float)
Diabetes	Presence of diabetes (N: No, Y: Yes, P: Pending)	Categorical (String/Object)

Fig. 1: Diabetes Dataset

Urea, Cr, HbA1c, Chol, TG, HDL, LDL, VLDL, and BMI. The dependent variable is the CLASS column, which indicates the presence or absence of diabetes (**Fig 2**).

The ”ID” and ”No_Pation” columns in the dataset serve as unique identifiers and do not hold any significant information for the prediction task of diabetes in logistic regression, therefore, they are typically excluded from the logistic regression analysis.



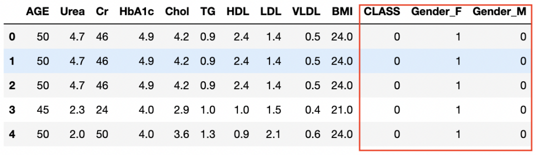
	ID	No_Pation	Gender	AGE	Urea	Cr	HbA1c	Chol	TG	HDL	LDL	VLDL	BMI	CLASS
0	502	17975	F	50	4.7	46	4.9	4.2	0.9	2.4	1.4	0.5	24.0	N
1	420	47975	F	50	4.7	46	4.9	4.2	0.9	2.4	1.4	0.5	24.0	N
2	680	87656	F	50	4.7	46	4.9	4.2	0.9	2.4	1.4	0.5	24.0	N
3	634	34224	F	45	2.3	24	4.0	2.9	1.0	1.0	1.5	0.4	21.0	N
4	721	34225	F	50	2.0	50	4.0	3.6	1.3	0.9	2.1	0.6	24.0	N

Fig. 2: Loaded Dataset

B. Data Exploration and Descriptive Statistics

This section focuses on conducting a descriptive analysis to enhance the understanding of the variables in the dataset. Summary statistics, such as mean, median, and standard deviation, are computed, and visualizations, including histograms and box plots, are utilized to explore the distributions of the variables and identify any outliers or patterns.

1) The unique values in the ”CLASS” column are checked, revealing three categories: ’N’ (No diabetes), ’P’ (Diabetes prediction unknown), and ’Y’ (Diabetes). For the purpose of analysis, the ’P’ cases are discarded, focusing only on the ’N’ and ’Y’ cases.

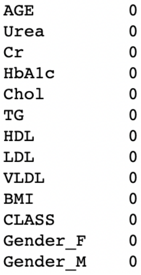


	AGE	Urea	Cr	HbA1c	Chol	TG	HDL	LDL	VLDL	BMI	CLASS	Gender_F	Gender_M
0	50	4.7	46	4.9	4.2	0.9	2.4	1.4	0.5	24.0	0	1	0
1	50	4.7	46	4.9	4.2	0.9	2.4	1.4	0.5	24.0	0	1	0
2	50	4.7	46	4.9	4.2	0.9	2.4	1.4	0.5	24.0	0	1	0
3	45	2.3	24	4.0	2.9	1.0	1.0	1.5	0.4	21.0	0	1	0
4	50	2.0	50	4.0	3.6	1.3	0.9	2.1	0.6	24.0	0	1	0

Fig. 3: Modified Dataset

2) To represent gender, the ”Gender” column is one-hot encoded into two separate columns, ”Gender_F” and ”Gender_M,” signifying the female and male genders, respectively where the presence of a particular gender is indicated by a binary value (1) and the absence by (0). This encoding technique is employed to convert categorical data (gender) into a numerical format that can be easily understood and processed by machine learning algorithms [2]. The original ”Gender” column is subsequently dropped from the dataset (**Fig 3**).

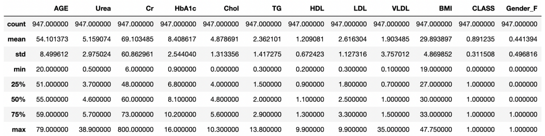
3) Null values are also examined, and no missing values are found across any of the columns (**Fig 4**).



```
AGE         0
Urea        0
Cr          0
HbA1c       0
Chol        0
TG          0
HDL         0
LDL         0
VLDL        0
BMI         0
CLASS       0
Gender_F    0
Gender_M    0
```

Fig. 4: Absence of Null Values in The Dataset

4) The describe() function generates summary statistics for each variable in the dataset, including count, mean, standard deviation, minimum, maximum, and quartile values [3]. These statistics provide valuable insights into the central tendency, spread, and range of values within the dataset (**Fig 5**).



	AGE	Urea	Cr	HbA1c	Chol	TG	HDL	LDL	VLDL	BMI	CLASS	Gender_F
count	947.000000	947.000000	947.000000	947.000000	947.000000	947.000000	947.000000	947.000000	947.000000	947.000000	947.000000	947.000000
mean	54.101373	5.159074	69.103485	8.408617	4.878691	2.362101	1.209081	2.616304	1.903485	29.893897	0.891235	0.441394
std	8.499612	2.975024	60.862961	2.544040	1.313356	1.417275	0.672423	1.127316	3.757012	4.869852	0.311508	0.496816
min	20.000000	0.500000	6.000000	0.900000	0.000000	0.300000	0.200000	0.300000	0.100000	19.000000	0.000000	0.000000
25%	51.000000	3.700000	48.000000	6.800000	4.000000	1.500000	0.900000	1.800000	0.700000	27.000000	1.000000	0.000000
50%	55.000000	4.600000	60.000000	8.100000	4.800000	2.000000	1.100000	2.500000	1.000000	30.000000	1.000000	0.000000
75%	59.000000	5.700000	73.000000	10.200000	5.600000	2.900000	1.300000	3.300000	1.500000	33.000000	1.000000	1.000000
max	79.000000	38.900000	800.000000	16.000000	10.300000	13.800000	9.900000	9.900000	35.000000	47.750000	1.000000	1.000000

Fig. 5: Dataset Description

- The mean value of ”HbA1c” indicates the average blood glucose level. The standard deviation suggests a moderate amount of variability in HbA1c values among the individuals. The minimum and maximum values represent the range of HbA1c values in the dataset. This range provides insight into the span of blood glucose levels observed.
- The quartile values give a sense of the distribution and central tendency of HbA1c levels among the patients. The median indicates that half of the patients have HbA1c levels below this value and the other half have levels above it.

5) In the data preprocessing step, outliers are removed from the dataset using two different approaches.

- The first approach involves removing extreme outliers using the interquartile range (IQR) method, Any value outside the range of Q1 - 1.5 * IQR to Q3 + 1.5 * IQR is considered an outlier and removed from the dataset [4]. This method helps to remove extreme outliers that could significantly affect the analysis (**Fig 6**).

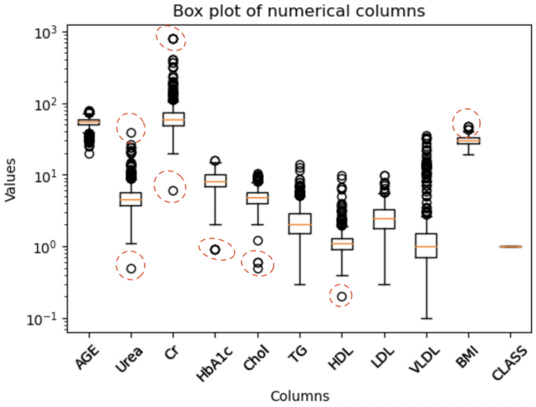


Fig. 6: After Removing Extreme Data Points Based on IQR

- Subsequently, a second approach is implemented to remove outliers that were more than three standard deviations away from the mean. This method relies on the assumption that values falling beyond three standard deviations from the mean are considered rare occurrences and likely to be outliers(**Fig 7**). The decision to use three standard deviations as the threshold to identify outliers is based on the empirical rule, also known as the 68-95-99.7 rule [5]. According to this rule, in a normal distribution, approximately 99.7% of the data falls within three standard deviations of the mean. Values beyond this range are considered rare and potential outliers.

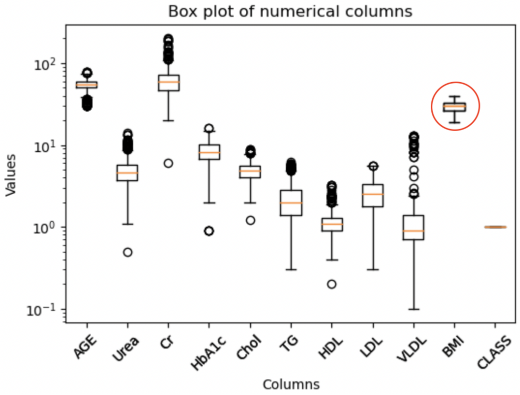


Fig. 7: Data Points Removed Based on Three Standard Deviation

By removing outliers, the study ensures that the dataset is more representative of the majority of observations and reduces the potential influence of extreme values on subsequent analysis and modeling.

6) **Data Visualization**

- A histogram of the variable 'AGE' stratified by the variable 'CLASS' is created to examine the distribution of age among different diabetes classes (**Fig 10**). The histogram displayed the frequency count of age values for each class of diabetes. The histogram reveals a higher frequency of older individuals(55-60 years old) in the Positive class compared to the Negative class. This observation suggests a potential association between age and diabetes, with older individuals being more likely to have the disease.

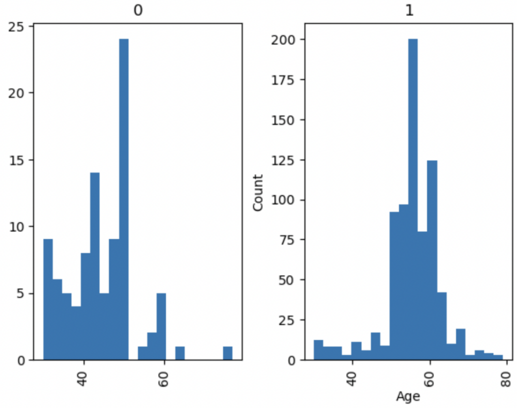


Fig. 8: Histogram of AGE by CLASS

- A scatterplot is created to visualize the relationship between the BMI (Body Mass Index) and Chol (Cholesterol) variables in the dataset (**Fig 9**). Each data point in the scatterplot represents a patient, and the color of the point represents the diabetes classification (CLASS).

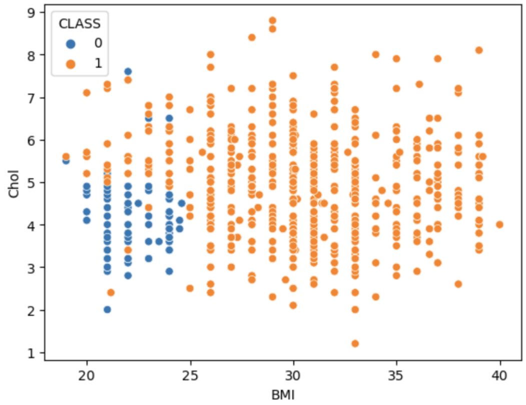


Fig. 9: Scatterplot of BMI vs Cholesterol

From the scatterplot, it can be observed that there is a scattered distribution of data points, suggesting a potential lack of a strong linear relationship between BMI and cholesterol levels.

Specifically, it is noticeable that a significant number of patients within the age range of 20-25 exhibit lower cholesterol levels and BMI values. This observation suggests that individuals in this age group, with lower cholesterol and BMI, have a reduced likelihood of being diagnosed with diabetes.

- The target variable, ”CLASS,” represents the diabetes classification of the patients. The dataset consists of 754 instances classified as diabetes-positive (1) and 94 instances classified as diabetes-negative (0) (**Fig 10**).

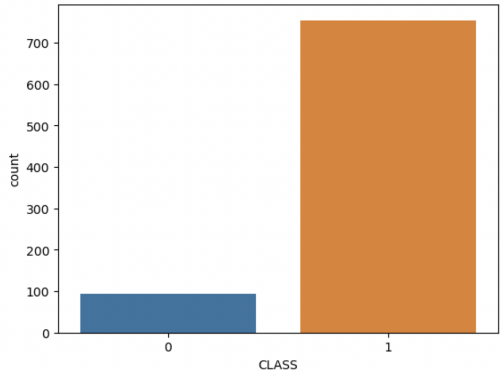
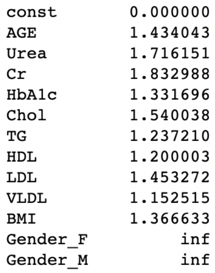


Fig. 10: Target Variable Count

7) Variance Inflation Factor (VIF) values are calculated to assess the presence of multicollinearity among the independent variables [6]. A VIF value of 5 or greater is typically considered indicative of significant multicollinearity (Fig 11). All variables have VIF values below the threshold of 5, suggesting that independent variables are not highly correlated with each other.



```
const        0.000000
AGE          1.434043
Urea         1.716151
Cr           1.832988
HbA1c        1.331696
Chol         1.540038
TG           1.237210
HDL          1.200003
LDL          1.453272
VLDL         1.152515
BMI          1.366633
Gender_F          inf
Gender_M          inf
```

Fig. 11: VIF Values of Independent Variables

To further examine the relationship between variables, the correlation coefficients are computed. A correlation coefficient of 0.7 or higher suggests a strong linear relationship (**Fig 12**)[7].

Based on the correlation coefficients, no variables exhibit a correlation coefficient of 0.7 or higher, indicating that there are no strong linear relationships between the variables. The selected independent variables are suitable for inclusion in the logistic regression model, as they do not exhibit multicollinearity.

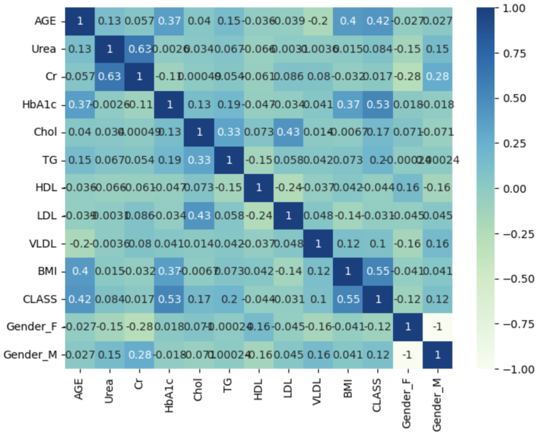
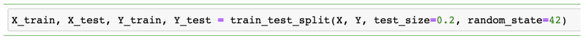


Fig. 12: Correlation Coefficients

C. *Model Building*

In the model-building phase, the dataset is split into independent features (X) and dependent feature (Y), representing the diabetes classification. The dataset is further divided into training and testing sets using the train_test_split function from the scikit-learn library [8], with a test size of 20% and a random state of 42 (for reproducibility) (**Fig 13**). The StandardScaler object from the scikit-learn library is used to scale the training and testing data to ensure standardized data. The independent features in the training set are transformed utilizing the scaler’s fit_transform method.



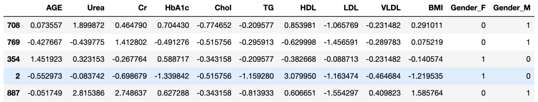
```
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, random_state=42)
```

Fig. 13: Data Split Criteria

The scaled independent features in the testing set have been transformed using the scaler previously fitted on the training set. The scaling process ensures consistency and allows the model to accurately predict unseen data.

The table presents a snapshot of the scaled independent features in the training set. The features, including AGE, Urea, Cr, HbA1c, Chol, TG, HDL, LDL, VLDL, and BMI, have been standardized using the StandardScaler.

The scaled features exhibit a mean of 0 and a standard deviation of 1, ensuring all variables are on a similar scale (**Fig 14**). It is crucial to scale the data when using models that are sensitive to the scale of the features, such as logistic regression.



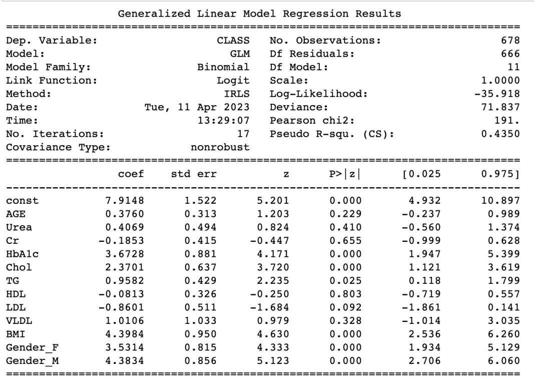
	AGE	Urea	Cr	HbA1c	Chol	TG	HDL	LDL	VLDL	BMI	Gender_F	Gender_M
708	0.073557	1.899872	0.464790	0.704430	-0.774652	-0.209577	0.853981	-1.065769	-0.231482	0.291011	0	1
769	-0.427667	-0.439775	1.412802	-0.491276	-0.515756	-0.295913	-0.629998	-1.456591	-0.289783	0.075219	0	1
354	1.451923	0.323153	-0.267764	0.588717	-0.343158	-0.209577	-0.382668	-0.088713	-0.231482	-0.140574	1	0
2	-0.552973	-0.083742	-0.698679	-1.339842	-0.515756	-1.159280	3.079950	-1.163474	-0.464684	-1.219535	1	0
887	-0.051749	2.815386	2.748637	0.627288	-0.343158	-0.813933	0.606651	-1.554297	0.409823	1.585764	0	1

Fig. 14: Scaled Independent Features

D. *Final Model Parameters and Assumptions*

- The logistic regression model is constructed to predict diabetes classification based on the selected features

[12]. The model is initially fitted using all available independent features (**Fig 15**).

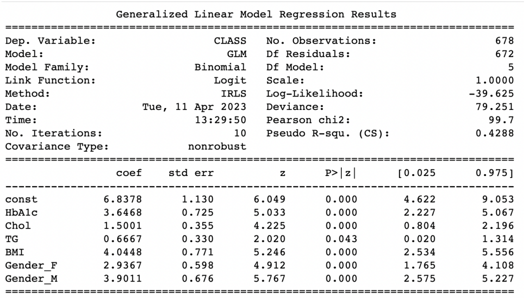


```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                  CLASS   No. Observations:                  678
Model:                            GLM   Df Residuals:                      666
Model Family:                Binomial   Df Model:                           11
Link Function:                  Logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -35.918
Date:                Tue, 11 Apr 2023   Deviance:                       71.837
Time:                        13:29:07   Pearson chi2:                     191.
No. Iterations:                    17   Pseudo R-squ. (CS):             0.4350
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          7.9148      1.522      5.201      0.000       4.932      10.897
AGE            0.3760      0.313      1.203      0.229      -0.237       0.989
Urea           0.4069      0.494      0.824      0.410      -0.560       1.374
Cr            -0.1853      0.415     -0.447      0.655      -0.999       0.628
HbA1c          3.6728      0.881      4.171      0.000       1.947       5.399
Chol           2.3701      0.637      3.720      0.000       1.121       3.619
TG             0.9582      0.429      2.235      0.025       0.118       1.799
HDL           -0.0813      0.326     -0.250      0.803      -0.719       0.557
LDL           -0.8601      0.511     -1.684      0.092      -1.861       0.141
VLDL           1.0106      1.033      0.979      0.328      -1.014       3.035
BMI            4.3984      0.950      4.630      0.000       2.536       6.260
Gender_F       3.5314      0.815      4.333      0.000       1.934       5.129
Gender_M       4.3834      0.856      5.123      0.000       2.706       6.060
==============================================================================
```

Fig. 15: The model with All Independent Features

The model summary shows that the constant term (const) had a coefficient of 7.9148 and a standard error of 1.522. The coefficient indicates that when all other independent variables are zero, the log odds of being classified as diabetes increase by a factor of exp(7.9148), or approximately 2704.59. The p-value for the constant term was less than 0.001, indicating its statistical significance.

Among the initial independent features, several variables are found to have p-values greater than 0.05, suggesting that they are not significantly associated with the diabetes classification. Therefore, these variables are removed from the model. The updated logistic regression model was then fitted using the remaining independent variables: HbA1c, Chol, TG, BMI, Gender_F, and Gender_M (**Fig 16**).



```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                  CLASS   No. Observations:                  678
Model:                            GLM   Df Residuals:                      672
Model Family:                Binomial   Df Model:                            5
Link Function:                  Logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -39.625
Date:                Tue, 11 Apr 2023   Deviance:                       79.251
Time:                        13:29:50   Pearson chi2:                     99.7
No. Iterations:                    10   Pseudo R-squ. (CS):             0.4288
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          6.8378      1.130      6.049      0.000       4.622       9.053
HbA1c          3.6468      0.725      5.033      0.000       2.227       5.067
Chol           1.5001      0.355      4.225      0.000       0.804       2.196
TG             0.6667      0.330      2.020      0.043       0.020       1.314
BMI            4.0448      0.771      5.246      0.000       2.534       5.556
Gender_F       2.9367      0.598      4.912      0.000       1.765       4.108
Gender_M       3.9011      0.676      5.767      0.000       2.575       5.227
==============================================================================
```

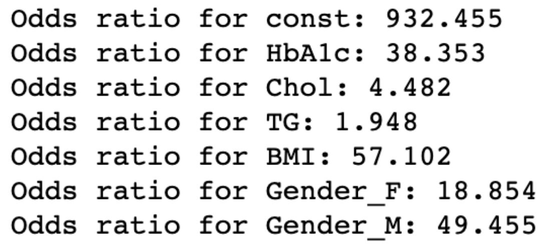
Fig. 16: The model with Selected Independent Features

The likelihood ratio test compares the fit of the current model to a null model (a model with no predictors) to determine if the predictors in the current model significantly improve the fit. The test statistic is based on the difference in the log-likelihood values between the two models [9].

The difference in log-likelihood values of 3.707 indicates that the inclusion of all 11 predictors in the first model slightly improves the fit compared to the second model with only 5 predictors. However, this improvement is not statistically significant based on the likelihood ratio test (p-value of 1.0000), suggesting that the reduced model is sufficient for explaining the relationship between the predictors and the response variable.

The coefficients of the remaining independent variables provide insights into their relationships with the diabetes classification. For each one-unit increase in HbA1c, the log odds of being classified as diabetes increase by a factor of exp(3.6468), or approximately 38.35 holding all other variables constant (**Fig 17**).

Similarly, for each one-unit increase in Chol, TG, BMI, Gender_F, and Gender_M, the log odds of being classified as diabetes increase by factors of exp(1.5001), exp(0.6667), exp(4.0448), exp(2.9367), and exp(3.9011), respectively.



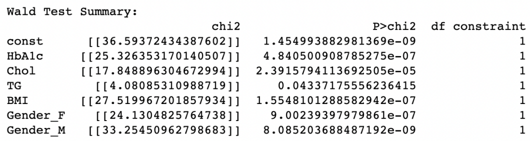
```
Odds ratio for const: 932.455
Odds ratio for HbA1c: 38.353
Odds ratio for Chol: 4.482
Odds ratio for TG: 1.948
Odds ratio for BMI: 57.102
Odds ratio for Gender_F: 18.854
Odds ratio for Gender_M: 49.455
```

Fig. 17: Odds Ratio of Selected Features

All remaining independent variables in the model are found to be statistically significant with p-values less than 0.05.

- **Wald Test Summary**: The Wald test is conducted to assess the significance of the coefficients in the logistic regression model. It is a statistical test that evaluates whether individual coefficients are significantly different from zero [10]. The test calculates a chi-square statistic and corresponding p-value for each coefficient.

  The results suggest that HbA1c, Chol, BMI, and gender variables (Gender_F and Gender_M) are significant predictors of the response variable (**Fig 18**).



```
Wald Test Summary:
                              chi2                  P>chi2  df constraint
const        [[36.59372434387602]]   1.454993882981369e-09              1
HbA1c       [[25.326353170140507]]    4.84050090878527e-07              1
Chol        [[17.848896304672994]]  2.3915794113692505e-05              1
TG            [[4.08085310988719]]     0.04337175556236415              1
BMI         [[27.519967201857934]]  1.5548101288582942e-07              1
Gender_F      [[24.1304825764738]]    9.00239397979861e-07              1
Gender_M     [[33.25450962798683]]   8.085203688487192e-09              1
```

Fig. 18: Wald Test

Based on the Wald test results, the interpretation of the significant predictor is done like below: - **BMI**: The chi-square statistic is 27.52 with a p-value of approximately 1.55e-07, indicating that the BMI coefficient is significantly different from zero. This suggests that BMI is a significant predictor of the response variable, and an increase in BMI is associated with higher odds of the outcome.

- **Chol**: The chi-square statistic is 17.85 with a p-value of approximately 2.39e-05, indicating that the Chol coefficient is significantly different from zero. This suggests that Chol is a significant predictor of the response variable, and an increase in Chol is associated with higher odds of the outcome.

IV. Model Performance

- The model is evaluated on the test dataset and the confusion matrix shows that out of 170 samples in the test dataset, 20 are correctly classified as negative (0), and 147 are correctly classified as positive (1), resulting in an overall accuracy of 98% (**Fig 19**).

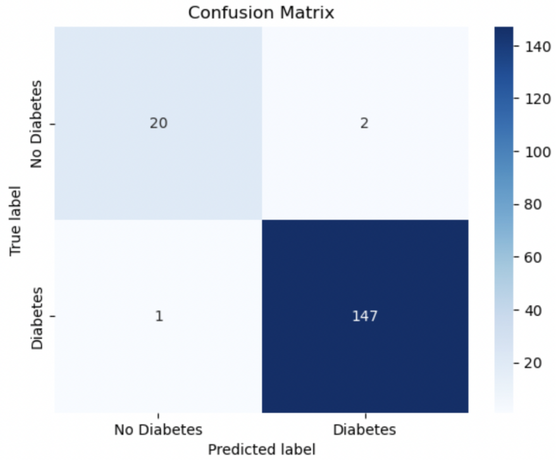
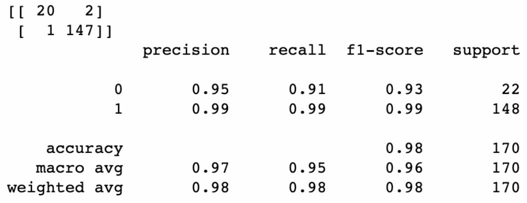


Fig. 19: Confusion Matrix

The classification report provides additional performance metrics [11]. The model achieved high precision (0.95 for class 0 and 0.99 for class 1), indicating a low rate of false positives. The recall values (0.91 for class 0 and 0.99 for class 1) indicate the model’s ability to correctly identify positive cases (**Fig 20**).



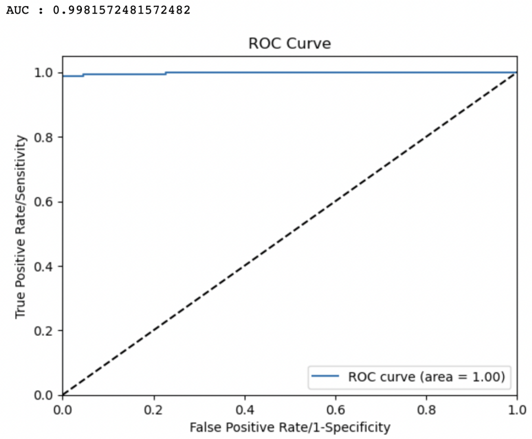
```
[[ 20   2]
 [  1 147]]
              precision    recall  f1-score   support

           0       0.95      0.91      0.93        22
           1       0.99      0.99      0.99       148

    accuracy                           0.98       170
   macro avg       0.97      0.95      0.96       170
weighted avg       0.98      0.98      0.98       170
```

Fig. 20: Evaluation Metrics

The f1-score, which combines precision and recall, is high for both classes (0.93 for class 0 and 0.99 for class 1), indicating overall good performance. These results suggest that the model performs well in classifying the test dataset.

- The ROC curve is a graphical representation of the true positive rate (sensitivity) against the false positive rate (1-specificity) for different threshold values and is commonly used to evaluate the performance of a binary classification model.

  The AUC represents the overall performance of the model, with a higher value indicating better discrimination between the positive and negative classes.

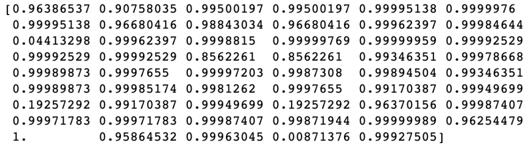


```
AUC : 0.9981572481572482
```

Fig. 21: Receiver operating characteristic (ROC) curve

In this study, the ROC curve and AUC are computed to assess the performance of the logistic regression model in predicting the outcome. The AUC value obtained is 0.998, indicating excellent discrimination between the classes (**Fig 21**). In the plotted ROC curve, the curve closely follows the top-left corner, further confirming the model’s excellent performance. The diagonal line from 0 to 1 represents the performance of a random classifier, which has an AUC of 0.5. Overall, the high AUC value and the shape of the ROC curve indicate that the logistic regression model is capable of accurately predicting the outcome and exhibits strong discriminatory power between the classes.

- In order to further analyze the model’s performance, the focus is on the ”P” cases from the original data set. The logistic regression model is used to obtain predicted probabilities for these cases (**Fig 22**).



```
[0.96386537 0.90758035 0.99500197 0.99500197 0.99995138 0.9999976
 0.99995138 0.96680416 0.98843034 0.96680416 0.99962397 0.99984644
 0.04413298 0.99962397 0.9998815  0.99999769 0.99999959 0.99992529
 0.99992529 0.99992529 0.8562261  0.8562261  0.99346351 0.99978668
 0.99989873 0.9997655  0.99997203 0.9987308  0.99894504 0.99346351
 0.99989873 0.99985174 0.9981262  0.9997655  0.99170387 0.99949699
 0.19257292 0.99170387 0.99949699 0.19257292 0.96370156 0.99987407
 0.99971783 0.99971783 0.99987407 0.99871944 0.99999989 0.96254479
 1.         0.95864532 0.99963045 0.00871376 0.99927505]
```

Fig. 22: The array of predicted probabilities for the ”P” cases

These values represent the predicted probabilities for each ”P” case as determined by the logistic regression model. The predicted probabilities range from very low values (e.g., 0.00871376) to high values close to 1 (e.g., 0.99999989).

To classify these probabilities into binary predictions, a threshold of 0.5 is set. The accuracy of the predicted probabilities for the ”P” cases is calculated by comparing the binary predictions to the actual labels. In this case, the accuracy is determined to be 92.45%, which means that the model’s predicted probabilities align with the actual outcomes.

The x-axis in the histogram represents the predicted probabilities of having diabetes, while the y-axis represents the frequency of ”P” cases falling within specific ranges of predicted probabilities.

The histogram indicates that a majority of ”P” cases have predicted probabilities between 0.07 and 0.10. This suggests that the model predicts with high confidence that most of these cases indeed have diabetes.

Additionally, a cluster below 0.02 implies that the model is highly confident in classifying these cases as non-diabetic. These individuals are likely to exhibit characteristics or features that strongly indicate the absence of diabetes according to the model’s learned patterns. (**Fig 23**).

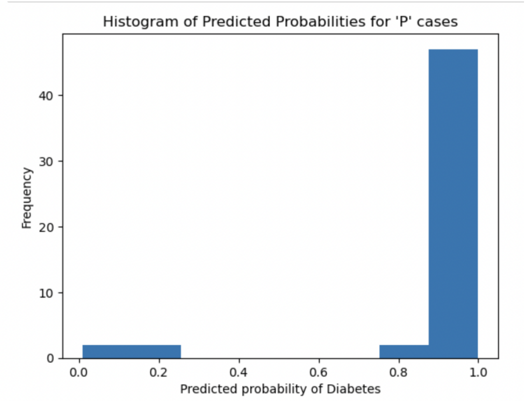


Fig. 23: Histogram of Predicted Probabilities in P Cases

The histogram demonstrates that the model can effectively distinguish between individuals with and without diabetes by assigning higher probabilities to cases known to have diabetes.

Hence, these findings suggest that the logistic regression model is a valuable tool for predicting the presence of diabetes in individuals, providing useful insights for diagnosis and treatment decisions.

V. Conclusion

In conclusion, logistic regression is applied to develop a predictive model for diabetes, considering various assumptions. These assumptions include having a binary or ordinal dependent variable, independent observations, minimal multicollinearity among independent variables, linearity of independent variables by removing outliers and transforming the data, and adequate sample size.

The model demonstrates good performance, as evidenced by high accuracy, recall, and a favorable confusion matrix. Additionally, odds ratios are calculated to interpret the effects of predictors on the odds of diabetes. The study’s findings emphasize the significance of logistic regression in accurately predicting diabetes and highlight the potential for improved diagnosis, treatment, and intervention strategies in this domain. The results underscore the importance of robust statistical methods in tackling the challenges associated with diabetes prediction and management.

References

[1] Priyanka Rajendra, Shahram Latifi, ”Prediction of diabetes using logistic regression and ensemble techniques”, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666990021000318/

[2] Amanda Fawcett, ”Data Science in 5 Minutes: What is One Hot Encoding?”, [Online]. Available: https://www.educative.io/blog/one-hot-encoding/

[3] pandas.DataFrame.describe [Online]. Available: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.describe.html/

[4] ”Identifying Outliers: IQR Method ” [Online]. Available: https://online.stat.psu.edu/stat200/lesson/3/3.2: :text=Any

[5] ”Empirical Rule: Definition, Formula, Example, How It’s Used” [Online]. Available: https://www.investopedia.com/terms/e/empirical-rule.asp/

[6] ”Variance Inflation Factor (VIF)” [Online]. Available: https://www.investopedia.com/terms/v/variance-inflation-factor.asp: :text=A

[7] ”Scatter Plots and Linear Correlation” [Online]. Available: https://k12.libretexts.org/Bookshelves/Mathematics/Statistics/02

[8] Scikit-learn [Online]. Available: https://scikit-learn.org/stable/supervised$_{l}earning.htmlsupervised-learning/$

[9] ”The Likelihood-Ratio Test” [Online]. Available: https://towardsdatascience.com/the-likelihood-ratio-test-463455b34de9/

[10] ”Wald Test: Definition, Examples, Running the Test” [Online]. Available: https://www.statisticshowto.com/wald-test//

[11] ”Classification Metrics Walkthrough: Logistic Regression with Accuracy, Precision, Recall, and ROC ” [Online]. Available: https://www.kdnuggets.com/2022/10/classification-metrics-walkthrough-logistic-regression-accuracy-precision-recall-roc.html: :text=in

[12] ”Simple Guide to Logistic Regression in R and Python ” [Online]. Available: https://www.analyticsvidhya.com/blog/2015/11/beginners-guide-on-logistic-regression-in-r//