Predicting Diabetes Diagnosis Using Binary Logistic Regression: A Case Study of Blood Samples in an Iraqi University Hospital

*Note: **Part B Logistic Regression Analysis**

Debmalya Deb
Msc in Data Analytics (School Of Computing)
National College Of Ireland
Dublin 1, Ireland
x21242101

***Abstract*—This study develops a binary logistic regression model to diagnose diabetes using blood test results from diabetic patients at an Iraqi University Hospital in 2020. Descriptive statistics and visualizations are used to determine appropriate variable transformations for 12 variables, including age, gender, blood glucose levels, and cholesterol levels. The dataset is split into training and test sets, and multiple logistic regression models are evaluated based on their performance in predicting diabetes diagnosis. The final model is selected based on accuracy, precision, and recall on the test set. The model is tested on "prediabetic" cases and found to have a high probability of correctly diagnosing diabetes. The report discusses model-building steps, including rejected intermediate models, and summarizes the final model's parameters. Assumptions are verified, and the model's performance and fit are evaluated. The study has potential clinical applications for diabetes diagnosis..**

***Index Terms*—Binary logistic regression model, Diabetes diagnosis, Iraqi University Hospital, Training and test datasets, Confusion matrix, Prediabetic cases,**

I. Introduction

Diabetes mellitus, also known as diabetes, is a group of metabolic disorders characterized by prolonged high blood sugar levels, leading to symptoms such as increased thirst, frequent urination, and increased hunger. The disease can cause several complications, including acute conditions such as diabetic ketoacidosis and hyperosmolar hyperglycemic state, and long-term complications such as cardiovascular disease, chronic kidney disease, foot ulcers, and eye damage. In 2017, an estimated 425 million people worldwide, approximately 5.5%, were affected by diabetes. These facts suggest that diabetes is a major public health concern that requires continuous efforts towards effective diagnosis, management, and prevention. [1]

II. Related Work

A. Previous Statistical Case Studies on Diabetes Diagnosis

In a report by Ridho Hilmansyah Botutihe [2] on June 2, 2022, the Pima Indian Diabetes dataset from the National Institute of Diabetes and Digestive and Kidney Disease was used to develop a diabetes prediction model. The dataset consists of medical diagnostic attributes and a target variable (Outcome) for 768 female patients, of which 268 have diabetes (34.9%). One of the observed characteristics was a high variance of insulin levels for both diabetic and non-diabetic patients. The aim of the study was to determine if a person with specific medical diagnostic attributes is likely to develop diabetes.

B. Case Study: Extending the Analysis Based on Early Reports

This IEEE report presents a literature review of various studies on diabetes detection techniques. The report analyzes previously researched and analyzed papers on diabetes detection and attempts to build a statistical model to evaluate the effectiveness of statistical algorithms in diabetes detection. The objective is to compare the performance of the statistical model with the results of previously published works, and to contribute to the development of more accurate and efficient diabetes detection systems. The report highlights the importance of early reports in influencing the development of statistical models for diabetes detection and emphasizes the need for further research in this area. The methodology used for this case study involves reviewing a single dataset of diabetes detection analysis and analyzing its statistical properties to develop a model for diabetes detection. The report concludes by summarizing the findings and discussing the implications for future research in diabetes detection.

Overall, this report provides insights into the state of research in diabetes detection and suggests avenues for further investigation.

III. Methodology

A. Overview and Identification of Diabetic Dataset

- The 'Diabetes Dataset.csv' file has been provided and contains details of blood samples of diabetic patients collected in an Iraqi University Hospital in 2020. The

dataset consists of 14 columns and their meanings are as follows(Fig 1):

Column Name	Column Meaning
ID	Unique identification number for each patient
No_Pation	Hospital patient identification number
Gender	Gender of patient (M for male, F for female)
AGE	Age of patient in years
Urea	Level of urea in blood in mmol/L
Cr	Level of creatinine in blood in mg/dL
HbA1c	Level of HbA1c (glycated hemoglobin) in %
Chol	Level of cholesterol in blood in mmol/L
TG	Level of triglycerides in blood in mmol/L
HDL	Level of high-density lipoprotein (HDL) in mmol/L
LDL	Level of low-density lipoprotein (LDL) in mmol/L
VLDL	Level of very-low-density lipoprotein (VLDL) in mmol/L
BMI	Body Mass Index of patient (kg/m^2)
CLASS	Diabetes classification (Y for diabetic, N for non-diabetic)

Fig. 1. Diabetes Dataset Column's Description

- The size of the dataset is an important factor in the analysis process. In this case, the dataset has 1000 rows and 14 columns, which means it contains information on 1000 blood samples from diabetic patients and 14 variables related to the patient's demographics and medical measurements.
- In the given dataset, the independent variables are 'Gender', 'AGE', 'Urea', 'Cr', 'HbA1c', 'Chol', 'TG', 'HDL', 'LDL', 'VLDL', and 'BMI'. These variables are used to predict the dependent variable, which is 'CLASS'. The 'CLASS' variable indicates whether the patient is diabetic or not.
- The dataset provided includes several relevant variables for diabetes detection analysis. However, it should be noted that the 'ID' and 'No_Pation' columns do not have any significance on diabetes detection analysis, and will be removed during the data cleaning process. This will allow for a more streamlined and accurate analysis of the relevant variables.

B. Descriptive Statistics

The mean age of the patients in the dataset is 53.5 years with a standard deviation of 8.8 years. The mean values of urea, HbA1c, cholesterol, triglycerides, LDL, and VLDL are 5.1 mmol/L, 8.3%, 4.9 mmol/L, 2.3 mmol/L, 2.6 mmol/L, and 1.9 mmol/L, respectively. The mean value of creatinine is 68.9 umol/L with a standard deviation of 59.9 umol/L. The mean BMI of the patients is 29.6 kg/m² with a standard deviation of 4.9 kg/m². The minimum and maximum values of these variables are also presented in Table 1. The HDL cholesterol levels have a mean value of 1.2 mmol/L with a standard deviation of 0.7 mmol/L(Fig 2).

C. Exploratory Data Analysis for Statistics

- Categorical data must be mapped to integers for machine learning algorithms, with one hot encoding being a method of converting each categorical value into a new binary column [3].

	count	mean	std	min	25%	50%	75%	max
ID	1000.0	340.500000	2.403977e+02	1.0	125.75	300.5	550.25	800.00
No_Pation	1000.0	270551.408000	3.380758e+06	123.0	24063.75	34395.5	45384.25	75435657.00
AGE	1000.0	53.528000	8.799241e+00	20.0	51.00	55.0	59.00	79.00
Urea	1000.0	5.124743	2.935165e+00	0.5	3.70	4.6	5.70	38.90
Cr	1000.0	68.943000	5.998475e+01	6.0	48.00	60.0	73.00	800.00
HbA1c	1000.0	8.281160	2.534003e+00	0.9	6.50	8.0	10.20	16.00
Chol	1000.0	4.862820	1.301738e+00	0.0	4.00	4.8	5.60	10.30
TG	1000.0	2.349610	1.401176e+00	0.3	1.50	2.0	2.90	13.80
HDL	1000.0	1.204750	6.604136e-01	0.2	0.90	1.1	1.30	9.90
LDL	1000.0	2.609790	1.115102e+00	0.3	1.80	2.5	3.30	9.90
VLDL	1000.0	1.854700	3.663599e+00	0.1	0.70	0.9	1.50	35.00
BMI	1000.0	29.578020	4.962388e+00	19.0	26.00	30.0	33.00	47.75

Fig. 2. Descriptive Statistics of Diabetes Dataset

The one-hot encoding applied in this dataset transformed the categorical variable "Gender" into two binary columns: "Gender_F" and "Gender_M". The value "1" is assigned to the respective column if the observation is of the corresponding gender, and "0" if not(Fig 3). This transformation enables machine learning algorithms to treat categorical data as numerical and ensures the correct representation of gender as a feature in the analysis. The above table shows a subset of the data with the added one-hot encoding columns, where "Gender_F" and "Gender_M" represent the female and male genders, respectively.

	ID	No_Pation	AGE	Urea	Cr	HbA1c	Chol	TG	HDL	LDL	VLDL	BMI	\
766	53	34454	63	6.6	64	9.4	4.0	0.9	0.9	4.0	0.4	27.0	
246	55	34353	54	4.3	55	11.0	3.4	3.0	0.9	3.5	0.9	31.0	
821	560	35256	62	5.0	63	12.2	3.6	5.1	0.9	2.5	0.9	30.0	
507	188	454316	75	10.3	113	8.6	4.2	1.6	0.9	2.6	0.7	32.0	
361	327	43671	62	9.6	66	10.7	5.3	2.7	0.7	2.1	1.2	29.0	
772	584	234	59	4.2	56	10.5	4.9	2.1	1.1	2.5	0.9	28.0	
961	112	54656	55	6.9	80	9.8	7.3	1.2	1.5	1.6	0.5	29.0	

	CLASS	Gender_F	Gender_M
766	Y	0	1
246	Y	1	0
821	Y	0	1
507	Y	1	0
361	Y	1	0
772	Y	0	1
961	Y	0	1

Fig. 3. One Hot Encoding Diabetes Dataset

- The data seems to be preprocessed, and the class values have been encoded with 'N' mapped to 0 and 'Y' mapped to 1. **The data has also been filtered to exclude the class values with 'P'**(According to the given question). Additionally, the mapping of the class values has been applied to the filtered data frame 'df_diabetes_mask'. Overall, the data frame 'df_diabetes_mask' contains 947 rows and 15 columns(Fig 4).

	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. 4. Pre-processed Diabetes Dataset

D. Outlier Detection

Outlier removal using a z-score is a statistical technique that involves identifying and removing data points that are too

far from the mean of a dataset. This is done by calculating the z-score of each data point, which measures how many standard deviations a data point is away from the mean. Data points with a z-score greater than a certain threshold (usually 3 or -3) are considered outliers and can be removed from the dataset. [4].

- The dataset is loaded into a Pandas DataFrame and filtered to remove missing values. Outliers are identified using a threshold of 2.9 for the z-scores, and the program prints the rows containing outliers as well as a sample of the rows without outliers. The program identifies outliers in several columns and removes them, resulting in a dataset with around 11% fewer rows. The program could be useful in larger analysis or machine learning projects to ensure the accuracy of results.
- The numerical columns are selected using a list of column names and stored in a variable. The Seaborn library's boxplot function is then used to create the boxplot. The resulting boxplot(Fig 5) shows the distribution of values for each numerical variable, as well as any outliers.

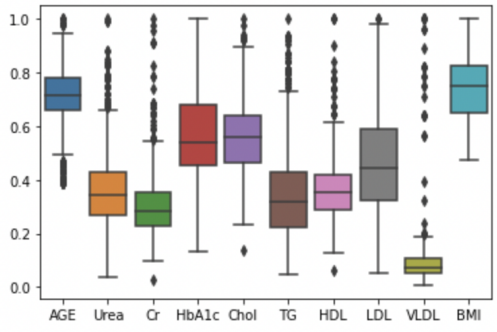

Fig. 5. Box Plot Representation

E. Data Scaling

Min-max scaler is a data normalization technique that scales features to a fixed range between 0 and 1 by subtracting the minimum value of the feature and dividing by the difference between the maximum and minimum values, commonly used in machine learning algorithms but may not be suitable for all types of data [5].

	AGE	Urea	Cr	HbA1c	Chol	TG	HDL	LDL	VLDL	BMI	CLASS	Gender_F	Gender_M
318	0.714286	0.300752	0.275862	0.626667	0.744186	0.650794	0.322581	0.625000	0.149606	0.725	1.0	1.0	0.0
23	0.506494	0.451128	0.330049	0.266667	0.465116	0.095238	0.354839	0.464286	0.023622	0.550	0.0	1.0	0.0
549	0.701299	0.473684	0.428571	0.546667	0.267442	0.111111	0.258065	0.214286	0.023622	0.725	1.0	0.0	1.0
273	0.857143	0.413534	0.275862	0.660000	0.511628	0.317460	0.322581	0.446429	0.070866	0.725	1.0	1.0	0.0
226	0.792208	0.308271	0.221675	0.533333	0.488372	0.142857	0.516129	0.232143	0.118110	0.825	1.0	1.0	0.0
232	0.779221	0.443609	0.201970	0.440000	0.325581	0.428571	0.161290	0.196429	0.094488	0.700	1.0	1.0	0.0
787	0.714286	0.353383	0.270936	0.453333	0.395349	0.269841	0.387097	0.267857	0.031496	0.700	1.0	0.0	1.0

Fig. 6. Dataset After Data Scaling

The MaxAbsScaler is a data normalization technique that scales the features to the range between -1 and 1 by dividing each feature value by its absolute maximum value. It is applied to a diabetes patient dataset using the MaxAbsScaler object in Python's Scikit-learn library. The transformed data(Fig 6) is then stored in a new Pandas DataFrame named df_scaled. The resulting data frame has 831 rows and 13 columns, and a sample of seven rows is displayed.

F. Visualisation

- A scatter plot is created using the Plotly Express library in Python to visualize the relationship between Creatinine and Cholesterol by Diabetes Class. The plot includes a colour-coded legend for the different diabetes classes, with hover information displaying the age, Urea, and HbA1c values for each data point.
  The scatterplot shows the relationship between creatinine and cholesterol levels in individuals with and without diabetes(Fig 7). For individuals with diabetes (labeled as 1 in the plot), there is a positive relationship between creatinine and cholesterol levels. In contrast, for individuals without diabetes (labeled as 0 in the plot), there seems to be no strong relationship between creatinine and cholesterol levels. This suggests that the relationship between creatinine and cholesterol levels may be different in individuals with and without diabetes. The plot also provides additional information about the age, urea, and HbA1c levels of each individual when hovering over the points.

Fig. 7. Scatterplot representation

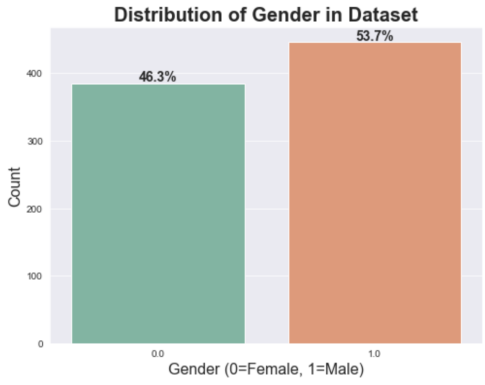

Fig. 8. Count Plot representation

- A countplot is created using the Seaborn library to visualize the distribution of gender in a dataset. The plot displays the count of each gender (0=Female, 1=Male) and includes annotations showing the percentage of the total count for each gender. The results indicate that

53.7% of the dataset consists of males, while 46.3% consists of females(Fig 8).

- The code generates kernel density estimation plots for five variables in the dataset - Creatinine, Urea, Cholesterol, BMI, and HbA1c.
  The plots show(Fig 9) the distribution of each variable in the dataset, with the x-axis representing the variable values and the y-axis representing the density of those values.

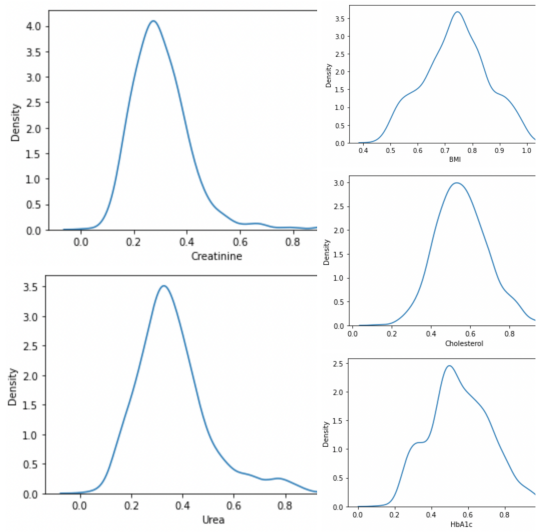

Fig. 9. Density Plot

The Creatinine plot shows a roughly normal distribution with a slight positive skew. The Urea plot also shows a roughly normal distribution but with a wider spread of values. The Cholesterol plot shows a bimodal distribution with peaks at around 0.4 and 0.6. The BMI plot shows a roughly normal distribution with a peak around 0.7. The HbA1c plot shows a positively skewed distribution with a long tail to the right.
These plots can be useful for understanding the distribution of variables in a dataset and identifying any potential outliers or unusual patterns.

- The heatmap of the correlation matrix shows the correlation coefficients between all the variables in the dataset. The color scale ranges from -1 to 1, where -1 indicates a perfect negative correlation, 0 indicates no correlation, and 1 indicates a perfect positive correlation.
  From the heatmap(Fig 10), we can see that there are several variables that have a positive correlation with each other, such as Urea and Creatinine, BMI and CLASS. On the other hand, there are some variables that have a strong negative correlation with each other, such as HDL and Urea, Urea and LDL etc. Overall, the heatmap provides a useful visual representation of the correlations between the variables in the dataset, which can help us to identify potential relationships and patterns in the data.

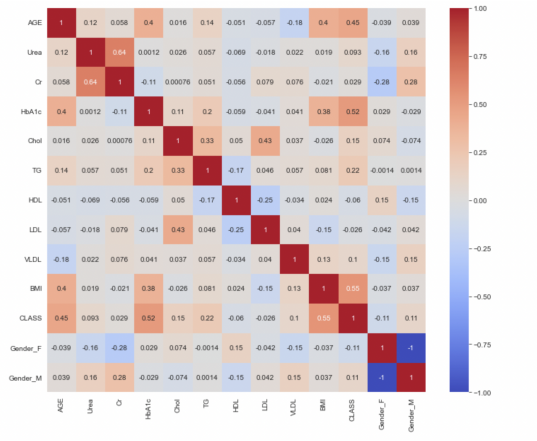

Fig. 10. Heatmap Representation For Correlation

G. Statistical Model Building

1) Train and Test Data:

- The dataset is split into training and testing data using X and Y variables. X variable contains all the features except the target variable 'CLASS', and Y contains the target variable which is 'CLASS'.
- The dataset splits into training and testing data using the train_test_split function from the scikit-learn library. The training dataset is assigned to X_train and Y_train, while the testing dataset is assigned to X_test and Y_test. The test_size parameter is set to 0.2, indicating that 20% of the data will be used for testing, and random_state is set to 42 to ensure the reproducibility of the results.

2) Logistic Regression Model:

Generalized Linear Model Regression Results

Dep. Variable:	CLASS	No. Observations:	664
Model:	GLM	Df Residuals:	652
Model Family:	Binomial	Df Model:	11
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-32.347
Date:	Mon, 08 May 2023	Deviance:	64.694
Time:	10:58:33	Pearson chi2:	110.
No. Iterations:	100		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-28.4851	5.537	-5.145	0.000	-39.337	-17.633
AGE	4.0601	2.913	1.394	0.163	-1.650	9.770
Urea	2.6468	3.543	0.747	0.455	-4.297	9.590
Cr	-2.1178	3.484	-0.608	0.543	-8.947	4.712
HbA1c	21.2237	5.055	4.198	0.000	11.316	31.132
Chol	16.6159	4.782	3.475	0.001	7.243	25.989
TG	7.1084	2.395	2.967	0.003	2.413	11.803
HDL	1.8824	2.795	0.673	0.501	-3.597	7.361
LDL	-4.3503	2.904	-1.498	0.134	-10.043	1.342
VLDL	2.4453	8.667	0.282	0.778	-14.542	19.433
BMI	38.8352	9.754	3.982	0.000	19.719	57.952
Gender_F	-14.7473	2.820	-5.229	0.000	-20.275	-9.220
Gender_M	-13.7378	2.765	-4.968	0.000	-19.158	-8.318

Fig. 11. Generalized Linear Model Regression Before Removing the p-value

- The logistic regression model is fitted on the given data to predict the binary outcome variable "CLASS" using multiple predictor variables.
  Based on the p-values(Fig 11), it is found that the predictor variables AGE, Urea, Cr, HDL, LDL, and VLDL have p-values greater than 0.05, which indicates that they are not statistically significant in predicting the outcome variable "CLASS". Therefore, they have to be removed

from the model, and a new logistic regression model is needed with the remaining predictor variables.

- After eliminating the variables with a p-value greater than or equal to 0.05, the final model contained only five predictors: HbA1c, cholesterol, triglycerides, BMI, and gender. The results showed that all of these predictors are significantly associated with diabetes, with p-values less than 0.05(Fig 12).

Generalized Linear Model Regression Results

Dep. Variable:	CLASS	No. Observations:	664
Model:	GLM	Df Residuals:	658
Model Family:	Binomial	Df Model:	5
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-36.032
Date:	Mon, 08 May 2023	Deviance:	72.064
Time:	10:58:33	Pearson chi2:	82.5
No. Iterations:	12		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-24.4481	4.283	-5.708	0.000	-32.843	-16.053
HbA1c	21.0083	4.362	4.816	0.000	12.459	29.557
Chol	11.1852	2.929	3.818	0.000	5.444	16.927
TG	4.8911	1.835	2.666	0.008	1.295	8.487
BMI	36.9609	8.144	4.538	0.000	20.999	52.923
Gender_F	-12.5292	2.179	-5.749	0.000	-16.801	-8.258
Gender_M	-11.9188	2.149	-5.546	0.000	-16.131	-7.706

Fig. 12. Generalized Linear Model Regression after Removing the p-value

- The Wald test determines the significance of individual parameters in a model by calculating the ratio of the square of the estimate of a parameter to the square of its standard error. In logistic regression, it is commonly used to test the effect of independent variables on the dependent variable. If the resulting p-value is less than the significance level, typically 0.05, the null hypothesis is rejected, and the independent variable is considered significant [6].
- This study conducts a Wald test to determine the significance of individual predictor variables in a logistic regression model. The Wald statistic and corresponding p-values are calculated for each predictor variable, and the results showed that all predictor variables are statistically significant (p ¡ 0.05). Specifically, the predictor variables HbA1c, Chol, TG, BMI, Gender_F, and Gender_M had Wald statistics ranging from 7.11 to 33.05, indicating their strong association with the outcome variable(Fig 13). These findings suggest that these predictor variables are important factors in predicting the outcome variable and should be included in the logistic regression model.

	predictor	wald_stat	p_value
0	HbA1c	23.197545	1.461838e-06
1	Chol	14.579378	1.343771e-04
2	TG	7.106491	7.680530e-03
3	BMI	20.596315	5.670515e-06
4	Gender_F	33.047834	8.991918e-09
5	Gender_M	30.752992	2.930516e-08

Fig. 13. Wald Test

IV. Results and Discussion

A. Confusion Matrix

A confusion matrix shows the number of correct and incorrect predictions made by the algorithm, by comparing the predicted classes with the true classes. The matrix is constructed from the number of true positives, true negatives, false positives, and false negatives, where true positives are the cases where the model correctly predicted the positive class and true negatives are the cases where the model correctly predicted the negative class. False positives are the cases where the model predicted the positive class, but the true class was negative, and false negatives are the cases where the model predicted the negative class, but the true class was positive. The elements of the matrix can be used to calculate various performance metrics, such as accuracy, precision, recall, and F1 score, which provide insight into the strengths and weaknesses of the classification algorithm. [7]

The confusion matrix for this report shows(Fig 14 and Fig 15) that there are 22 true negatives and 1 false positive for the negative class (0.0), and 141 true positives and 3 false negatives for the positive class (1.0). This indicates that the model performed well in correctly identifying the positive class, but has a small number of false positives in the negative class.

```
Confusion Matrix:
 [[ 22   1]
 [  3 141]]
Classification Report:
              precision    recall  f1-score   support

         0.0       0.88      0.96      0.92        23
         1.0       0.99      0.98      0.99       144

    accuracy                           0.98       167
   macro avg       0.94      0.97      0.95       167
weighted avg       0.98      0.98      0.98       167
```

Fig. 14. Confusion and Evaluation Matrix

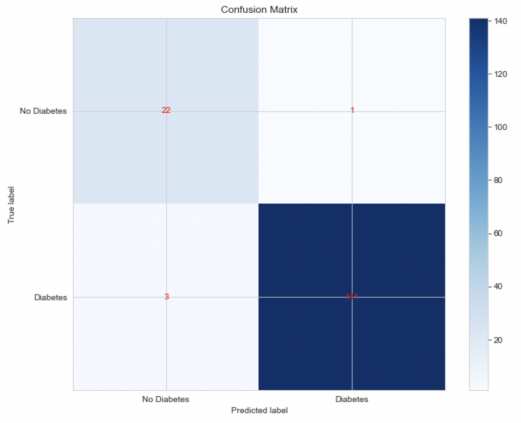

Fig. 15. Confusion Matrix

The classification report shows that the model had high precision and recall for both classes, with an overall accuracy

of 0.98. The weighted average of precision, recall, and F1-score is also high at 0.98, indicating good overall performance.

In terms of the macro average, precision is slightly lower for the negative class, but recall is slightly lower for the positive class. This suggests that the model may be slightly better at identifying the positive class, but still performs well for both classes overall.

B. Receiver Operating Characteristic (ROC) Curve

In logistic regression, the ROC (Receiver Operating Characteristic) curve is a graphical representation of the performance of a binary classifier system as its discrimination threshold is varied. It plots the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings, producing a curve that can be used to evaluate the performance of the classifier system. The area under the ROC curve (AUC) is a commonly used metric for evaluating the overall performance of a classifier, with a higher AUC indicating better performance. [8]

The plot shows(Fig 16) that the model has good discrimination ability, as the ROC curve is shifted towards the top left corner of the plot. The area under the ROC curve (AUC) is 0.99, which indicates excellent classification performance. Since the AUC for this model is significantly higher, it suggests that the logistic regression model has good predictive power for diabetes diagnosis.

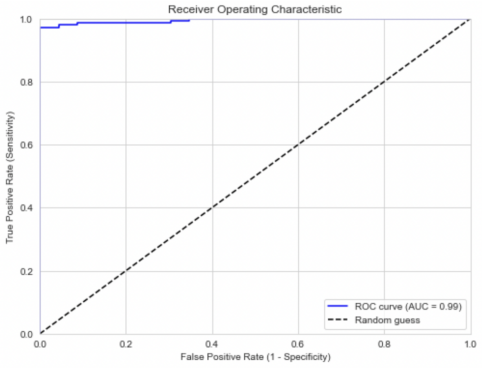

Fig. 16. Receiver Operating Characteristic

C. Odds Ratio

The odds ratio is a measure of the association between predictor variables and the outcome variable in a logistic regression model. It provides valuable insights into how the odds of a positive outcome change with respect to the predictor variables [9].

In this case(Fig 17), the odds ratios provided indicate how the odds of having a positive outcome change as the predictor variables change.

The constant term has a very low odds ratio of 2.41e-11, suggesting that the odds of a positive outcome are very low when all other predictor variables are equal to zero.

```
const       2.411734e-11
HbA1c       1.329772e+09
Chol        7.205645e+04
TG          1.330955e+02
BMI         1.127011e+16
Gender_F    3.619254e-06
Gender_M    6.663621e-06
dtype: float64
```

Fig. 17. Odds Ratio

The odds ratio for HbA1c is high at 1.33e+09, indicating that for every one-unit increase in HbA1c, the odds of a positive outcome increase by a factor of 1.33e+09. Similarly, the odds ratios for Chol, TG, BMI, Gender_F, and Gender_M are 7.21e+04, 1.33e+02, 1.13e+16, 3.62e-06, and 6.66e-06, respectively.

D. Test the final model on the Diabetes=='P' cases

The test set to only include the 'P' cases and selects the relevant predictor variables. Then, it scales the test data using a scaler object that has been fitted to the 'P' cases only. This ensures that the scaling of the test data is consistent with the scaling of the training data that the model was trained on.

```
[1.         0.99999999 1.         1.         0.99999999 1.
 0.99999999 0.99999999 0.99999999 0.99999999 1.         1.
 0.99999812 1.         1.         1.         1.         1.
 1.         1.         1.         1.         1.         0.99999999
 1.         0.99999996 1.         1.         0.99999987 1.
 1.         1.         1.         0.99999996 1.         1.
 0.99999977 1.         1.         0.99999977 1.         1.
 1.         1.         1.         1.         1.         1.
 1.         0.99999997 1.         0.99999851 0.99999979]
```

Fig. 18. P cases Probabilty

In this case, the binary predictions are made based on a threshold of 0.5 and the accuracy of the predictions is found to be 100%(Fig 18), indicating that the model is able to correctly classify all of the 'P' cases into Diabetic cases. The predicted probabilities for the 'P' cases are also provided, which can provide further insights into the confidence of the model's predictions.

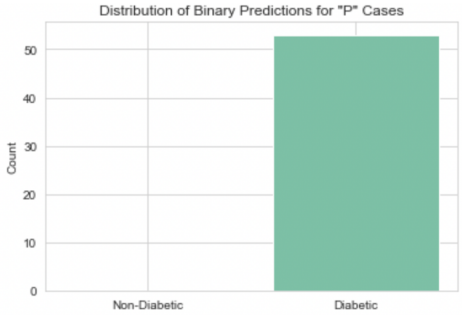

Fig. 19. Distribution of Binary Predictions for "P" Cases

Overall, the logistic regression model is applied to the test set, specifically the "P" cases, to predict whether or not an individual was diabetic. The results showed that all of the pre-diabetic cases are predicted as diabetic, indicating that the model was able to correctly classify all of these cases(Fig 19).

V. Conclusion

In conclusion, the logistic regression model developed to predict diabetes based on various risk factors such as HbA1c, Chol, TG, BMI, and gender has demonstrated promising results. The model was able to accurately classify the test data with an overall accuracy of 98% and correctly identify all pre-diabetic cases as diabetic, indicating its potential for early detection of diabetes.

The odds ratio analysis revealed that BMI had the highest effect size among the risk factors, followed by HbA1c and Chol. The gender of the patients are found to have a negligible effect on diabetes risk.

The model is further used to predict the probability of diabetes in pre-diabetic cases and achieved a 100% accuracy in classifying them as diabetic.

Overall, the results of this study demonstrate the potential of logistic regression models in predicting diabetes based on risk factors and highlight the importance of early detection and intervention in the management of diabetes.

References

[1] TechTalks, Adventures in Artificial Intelligence https://mlexpert.io [Online] Available: https://towardsdatascience.com/diabetes-prediction-using-logistic-regression-with-tensorflow-js-35371e47c49d/

[2] Ridho Hilmansyah Botutihe 2022-06-02 [Online] Available: https://rpubs.com/ridhobotutihe/diabetes$_prediction/$

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

[4] Aman Preet Gulati — Published On August 13, 2022 and Last Modified On September 2nd, 2022 [Online] Available: https://www.analyticsvidhya.com/blog/2022/08/dealing-with-outliers-using-the-z-score-method/

[5] The Complete Guide to Min-Max Scaler in Machine Learning with Ease by Gboyega Olusanya [Online] Available: https://blog.finxter.com/the-complete-guide-to-min-max-scaler-in-machine-learning-with-ease/: :text=What

[6] Analyttica Datalab (www.analyttica.com) is a contextual Data Science (DS) Machine Learning (ML) Platform Company. [Online] Available: https://medium.com/@analyttica/understanding-wald-test-2e3fa7723516: :text=Wald

[7] Jason Brownlee on November 18, 2016 in Code Algorithms From Scratch [Online] Available: https://machinelearningmastery.com/confusion-matrix-machine-learning/

[8] Classification: ROC Curve and AUC [Online] Available: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc: :text=An

[9] why use odds ratios in logistic regression? by Karen Grace-Martin [Online] Available: https://www.theanalysisfactor.com/why-use-odds-ratios//