**Predicting Diabetes Using Machine Learning Models: A Data-Driven Clinical Approach**

**Sivamugunthan ashok 1RVU22BBA091**

**1.Introduction**

Diabetes is one of the most fastest growing disease globally due to generation gene transfer else consumption of more natural sugar content products. This disease is characteristics by persistent hyperglycemia which often leads to critical issues like cardiovascular diseases, kidney failure, lower limb amputation, and vision loss. In 2014, on the authority of world health organization (WHO), roughly 422 million adults were living with diabetes, in 2024, 590 million people aged between (20 -75) have diabetes which is increase by 168 million which is 40 percentage. This disease is expected to grow by 1.5 million death each year leading to seventh leading cause of death worldwide (WHO, 2024).

While early diagnosis provides an opportunity to identify and prevent the disease earlier, it often leads to high resource and time consumption to treat the disease in traditional way. This growing number provides an opportunity to create a data driven decision. The integration of machine learning in health care sector promotes accurate and reliable path to identify and cure the disease. Logistic regression analysis is powerful tool of machine learning algorithm, which can be used to predict the binary classification-based outcome like categorical variable—such as diabetic vs. non-diabetic. It justifies and understand the cause and provides prediction according to the numerical health result report of the people.

This study considers factors such as pregnancies, glucose, blood pressure, skin thickness, Insulin, body mass index, diabetes pedigree function, Age to predict whether the person have diabetes. By evaluating the model accuracy and prediction performance, this research gives its potential baseline predictive model to the health care industry and contribute to the emerging artificial intelligence field.

**2.Objective**

This study aims to predict the diabetes through the people medical report results such pregnancies, glucose, blood pressure, skin thickness, Insulin, body mass index, diabetes pedigree function, age through the machine learning model. The specific objective of this research is to:

1. To know the relationship between various medical indicators (such as glucose level, BMI, insulin, age, etc.)
2. To know the key characteristics of the diabetes dataset to uncover the underlying patterns, relationships, insights.
3. To know the effectiveness of logistic regression as a predictive model for identifying diabetic and non-diabetic individuals based on people medical report.

**3. Results**

3.1 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) were conducted to understand the structure and quality of the dataset to predict the diabetes prediction. The original dataset was a structure table which contains rows representing the patients record and columns corresponding to various health indicators such as glucose levels, insulin, BMI, age, and a target variable labeled "Outcome" indicating diabetic or non-diabetic status. Data inspection conducted to conform the type of column. All the columns were numerical with no null values detected.

Table 1.1: Data information: *columns, non-null counts and data types*

|  |  |  |
| --- | --- | --- |
| **Column** | **Non-Null Count** | **Dtype** |
| Pregnancies | 768 non-null | int64 |
| Glucose | 768 non-null | int64 |
| BloodPressure | 768 non-null | int64 |
| SkinThickness | 768 non-null | int64 |
| Insulin | 768 non-null | int64 |
| BMI | 768 non-null | float64 |
| DiabetesPedigreeFunction | 768 non-null | float64 |
| Age | 768 non-null | int64 |
| Outcome | 768 non-null | int64 |

The independent and dependent variable were named by x and y. Additionally, a constant term was added to independent variables. This constant is representing the outcome variable when the independent variable is 0. This EDA process ensured the dataset is clean, well structed and can be used for modeling. The dataset contains 768 observations and 9 variables such as Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age, and the target variable Outcome named diabetes.

Initially the logistic regression with all the predictors had a Pseudo R-squared value of 0.272 which shows that the power of explanatory variable is moderate. The predictors are conformed as significant by checking its p value is lesser than 0.05. The predictors named insulin skinthickness and age is not significant means it doesn't have good relationship with the outcome variable. Table 1.2 shows the logistic regression summary.

Table 1.2: Logistic regression summary

|  |  |  |  |
| --- | --- | --- | --- |
| Metrics 1 | Values 1 | Metrics 2 | Values 2 |
| Model: | Logit | Method: | MLE |
| Dependent Variable: | diabetes | Pseudo R-squared: | 0.272 |
| Date: | 2025-04-11 19:02 | AIC: | 741.4454 |
| No. Observations: | 768 | BIC: | 783.2395 |
| Df Model: | 8 | Log-Likelihood: | -361.72 |
| Df Residuals: | 759 | LL-Null: | -496.74 |
| Converged: | 1.0000 | LLR p-value: | 9.6516e-54 |
| No. Iterations: | 6.0000 | Scale: | 1.0000 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Coef.** | **Std. Err.** | **z** | **P>|z|** | **[0.025** | **0.975]** |
| const | -8.4047 | 0.7166 | -11.7280 | 0.0000 | -9.8093 | -7.0001 |
| Pregnancies | 0.1232 | 0.0321 | 3.8401 | 0.0001 | 0.0603 | 0.1861 |
| Glucose | 0.0352 | 0.0037 | 9.4814 | 0.0000 | 0.0279 | 0.0424 |
| BloodPressure | -0.0133 | 0.0052 | -2.5404 | 0.0111 | -0.0236 | -0.0030 |
| SkinThickness | 0.0006 | 0.0069 | 0.0897 | 0.9285 | -0.0129 | 0.0141 |
| Insulin | -0.0012 | 0.0009 | -1.3223 | 0.1861 | -0.0030 | 0.00006 |
| BMI | 0.0897 | 0.0151 | 5.9453 | 0.0000 | 0.0601 | 0.1193 |
| DiabetesPedigreeFunction | 0.9452 | 0.2991 | 3.1596 | 0.0016 | 0.3589 | 1.5315 |
| Age | 0.0149 | 0.0093 | 1.5929 | 0.1112 | -0.0034 | 0.0332 |

To improve the model performance, variables with less than 0.05 p value is removed, which is not significant to predict the outcome. The final logistic regression table contains all significant variable which has good potential to predict the outcome variable.

Table 1.3: Final Logistic regression summary

|  |  |  |  |
| --- | --- | --- | --- |
| Metrics 1 | Values 1 | Metrics 2 | Values 2 |
| Model | Logit | Method | MLE |
| Dependent Variable | diabetes | Pseudo R-squared | 0.272 |
| Date | 2025-04-11 19:02 | AIC | 741.4454 |
| No. Observations | 768 | BIC | 783.2395 |
| Df Model | 8 | Log-Likelihood | -361.72 |
| Df Residuals | 759 | LL-Null | -496.74 |
| Converged | 1.0000 | LLR p-value | 9.6516e-54 |
| No. Iterations | 6.0000 | Scale | 1.0000 |

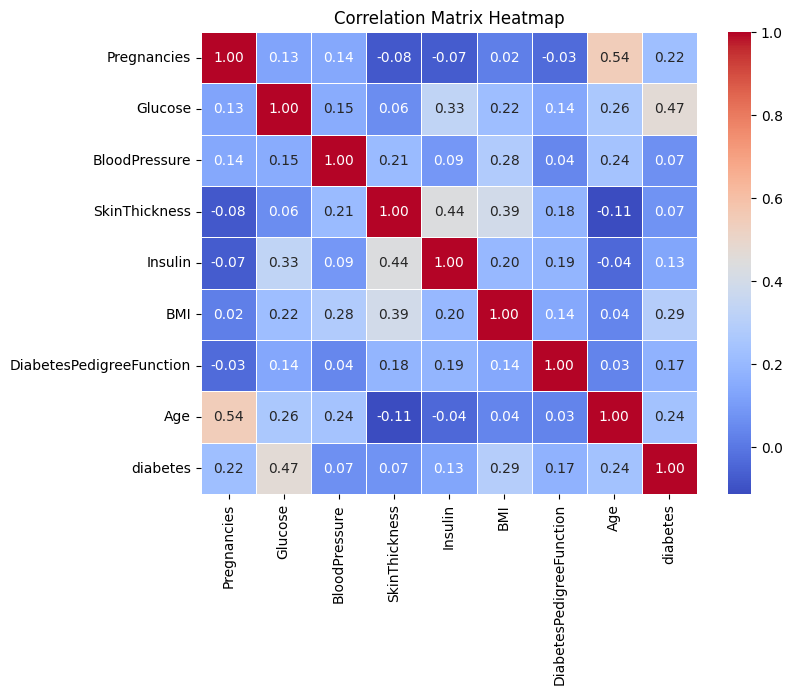
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variable | Coef. | Std.Err. | z | P>|z| | [0.025 | 0.975] |
| const | -7.9550 | 0.6758 | -11.7708 | 0.0000 | -9.2795 | -6.6304 |
| Pregnancies | 0.1535 | 0.0278 | 5.5143 | 0.0000 | 0.0989 | 0.2080 |
| Glucose | 0.0347 | 0.0034 | 10.2130 | 0.0000 | 0.0280 | 0.0413 |
| BloodPressure | -0.0120 | 0.0050 | -2.3868 | 0.0170 | -0.0219 | -0.0021 |
| BMI | 0.0848 | 0.0141 | 6.0059 | 0.0000 | 0.0571 | 0.1125 |
| DiabetesPedigreeFunction | 0.9106 | 0.2940 | 3.0971 | 0.0020 | 0.3343 | 1.4869 |

The model demonstrates a moderate explanatory power with a pseudo-R-squared value of 0.272, which indicates that 27.2 percent of the variance in dependent variable (Diabetes) is explained by the independent variables such as pregnancies, Glucose, bloodpressure, BMI, Diabetespedigreefunction which is acceptable for predictive modeling in health sciences. The model identified several significant factors among these the variable glucose has exhibited the highest Z score value (z = 10.2130), so it plays a crucial role in predicting the diabetes.

**3.2 Correlation Matrix (Heatmap)**

The correlation matrix shows key relationship between the health indicators. In the table 1.4 we can see Glucose (0.47), BMI (0.29) and Age (0.24) has the strongest positive correlation with diabetes. These indicators have the high potential to predict the outcome variable (diabetes).

Table 1.4: Correlation Matrix

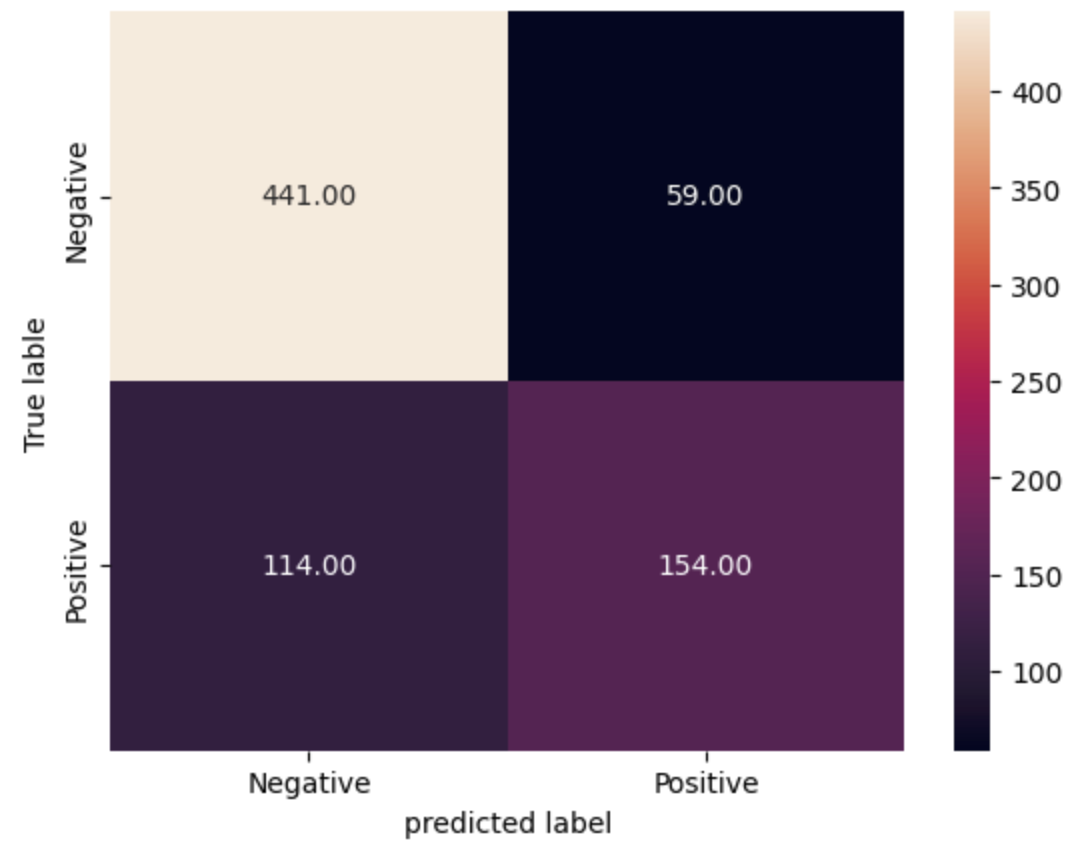


Medical indicators Variables such as Insulin (0.13), DiabetesPedigreeFunction (0.17), and BloodPressure (0.07) has weaker positive correlation with the outcome variable. SkinThickness shows very week negative correlation of (–0.07). Moreover, correlation matrix shows the relationship between the explanatory variables and outcome variable. These initial insights provide a primary value into assessing the linear associations within the dataset. It is important to note that further analysis is required to understand the underling factors to know predictive power of the independent variable.

3.3 Confusion Matrix

Confusion Matrix is a table that compares the predicted value with the actual value to know the predicting accuracy of the model. Specifically, it categorizes predictions into four components such as True Positive (TP), False Positive (FP), False Negative (FN), True Negative (TN). In table 1.5 True positive shows that 436 instances were correctly predicted as Negative Diabetes. False Positive shows that 64 instances were incorrectly predicted as positive when they were actually negative Diabetes. False negative show that 113 instances were incorrectly predicted as Negative Diabetes when they were actually positive Diabetes. True negative shoes that 155 instances were correctly predicted as positive Diabetes.

Table 1.5 Confusion Matrix



3.4 Classification report

The classification report provides a detailed evaluation of model performance between two classes which is labeled as 0 and 1. The class 0, the model shows the strong precision (0.79) and high recall (0.88), which results in a robust F1-score of 0.84, with 500 actual instances. While the positive predictions for class 1 have reasonable accuracy. A Macro-avg F1 score of 0.74 and a weighted F1 score of 0.77 continue to highlight the overall performance of the model. The observed performance differences between the two classes are probably due to this imbalance, with the model being skewed compared to the majority class (class 0).

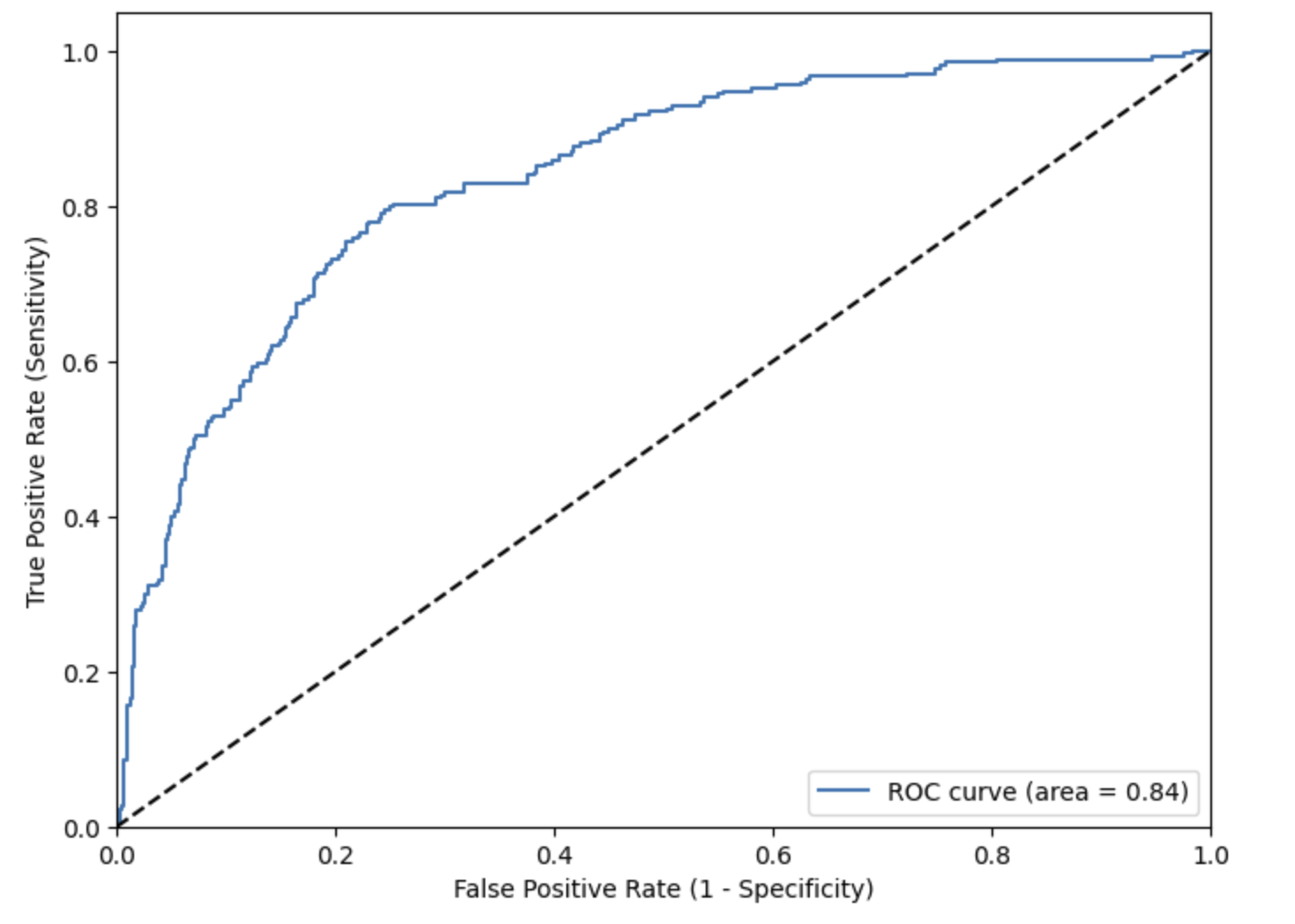
Table 1.6 Classification report

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class | Precision | Recall | F1-Score | Support |
| 0 | 0.79 | 0.88 | 0.84 | 500 |
| 1 | 0.72 | 0.57 | 0.64 | 268 |
| accuracy |  |  | 0.77 | 768 |
| macro avg | 0.76 | 0.73 | 0.74 | 768 |
| weighted avg | 0.77 | 0.77 | 0.77 | 768 |

3.5 Receiver operator curve (ROC) & Area under the curve (AUC)

These two curve were used to distinguish between two class 1 and 0 which indicated whether the patient have diabetes or not. The receiver operator curve in the diagram 1.7 shows that how the model’s True positive rate (Sensitivity) changing according to the False Positive rate which is 1-specificity) at diverse threshold levels. The diagonal line represents the random guessing with (AUC) value of 0.84 which shows that **84% chance** that the model will rank a randomly chosen positive instance higher than a randomly chosen negative instance. Area under the curve (AUC) shows the model performance is well because the AUC value is close to 1.

Diagram 1.7: Receiver operator curve (ROC) & Area under the curve (AUC)



**4. Conclusion**

When it comes to predicting diabetes using structured health data, logistics regression models show good fundamental functions. With an accuracy of 75.32% and ROC-AUC of 81.47%, this model distinguishes between people with diabetes and those without diabetes. The model shows a satisfactory performance, with a pseudo-R-squared value of 0.272 which indicated a moderate but meaningful explanatory variable capability to predict the outcome. Medical predictors (Independent variables) such as glucose level, BMI, diabetes pedigree function, and the number of pregnancies were statically significant which has a good impact on the predicting the likelihood of having diabetes. Among all the predictors glucose level has the higher significant score which plays the crucial part in predicting the diabetes. The finding of this study not only contributing to the logistic regression in medical diagnosis but also gives valuable insights that creates a huge convenience and reliability in finding and preventing diabetes. Overall, this study helps in understanding the diabetes risk factors and showcase how data driven approaches help in making accurate decision making.

**5. Discussion and Implication**

The findings from the logistic regression model provides valuable insights into factors that is leading to diabetes. This study shows that variables such as glucose level, number of pregnancies, BMI, and diabetes pedigree function were found that statistically significant predictors which aligns with the existing medical literature report by [(Priyanka Rajendra)](https://www.sciencedirect.com/science/article/pii/S2666990021000318) in her study. They have used PIMA Indian Diabetes dataset and identify almost similar results produced by this study. In this model, glucose level shows the highest z-score, which plays a critical role in predicting the likelihood of diabetes, as supported by [Mousa et al. (2024)](https://www.researchgate.net/publication/377406725_Prediction_of_Type_2_Diabetes_using_logistic_regression_techniques_Prediction_of_Type_2_Diabetes), who produced about 82% prediction accuracy using logistic regression. Similarly, BMI and diabetes pedigree function are not significant in this model which shows the impact of both lifestyle and genetic factors, in the findings of [Dwivedi et al. (2023)](https://pmc.ncbi.nlm.nih.gov/articles/PMC10465890/), who shown these variables as not significant in their medical study of likelihood of predicting diabetes. Our model had a pseudo-R-squared value of 0.272, which means it indicates a good explanation of diabetes by the medical predictors. The p-value was less than 0.05 which is meaning for this model is statistically significant. These results are almost similar to the study Zhou et al. (2021). Overall, our study shows the similar kind of results that the other researchers have found. Also, all study shows that logistic regression is a good tool to predict diabetes.