Data visualisation and an ensemble approach for the prediction of diabetes using logistic regression

"Exploring the Diabetes Prediction Frontier: A Journey through Logistic Regression, Ensemble Techniques, and Visual Storytelling."

ABSTRACT

The Pima Indian community in Arizona has exhibited alarmingly high rates of type 2 diabetes, prompting intensive research into predictive modeling to identify at-risk individuals. This study aims to predict the likelihood of diabetes in Pima Indian women using logistic regression analysis, incorporating key demographic and clinical variables. The data used in this study were sourced from Kaggle as secondary data to predict diabetes, including age, BMI, family history of diabetes, and glucose levels. Logistic regression models were constructed to assess the association between these variables and the binary outcome of diabetes status. The predictive performance of the models was evaluated using metrics such as sensitivity, specificity, and area under the receiver operating characteristic curve. The results indicated a significant correlation between age, BMI, and family history of diabetes with the likelihood of diabetes in Pima Indian women. The logistic regression model demonstrated good predictive accuracy, with implications for targeted preventive interventions. This research contributes to the growing body of literature on diabetes prediction in high-risk populations, emphasizing the importance of tailored approaches for specific ethnic groups such as the Pima Indians. Further refinement of predictive models and exploration of additional risk factors are warranted to enhance the precision of diabetes risk assessment in this vulnerable population.

1 INTRODUCTION

1.1 Background

Diabetes Mellitus has become a common health problem nowadays, which would affect people and lead to various disablements like cardio vascular disease, visual impairments, leg amputation and renal failure if diagnosis is not done in the right time. Diabetes can affect people due to the lack of insulin in the blood. Insulin is a natural hormone secreted by the pancreas, which acts as a key to unlock the body cells so that sugar, starch and food molecules can be absorbed and hence be utilized by the cells to generate energy required for daily life. Insulin deficiency is due to either of the two conditions. First is when the pancreas does not produce insulin at all. This leads to type I diabetes mellitus (T1DM) which is usually found by birth. Second state is when the body does not respond correctly to the insulin produced by the pancreas and hence the glucose that is consumed by the person is locked inside the blood instead of entering into the cells of the body. This ineffective insulin leads to type II diabetes mellitus (T2DM). Among these, type I diabetes is usually diagnosed in children and type II is the most common form which affects adults.

1.1 Diabetes-A Global Threat

The International Diabetes Federation has estimated an alarming rise in the number of diabetics by the year 2030, A sharp rise in diabetics has been observed in Asian region with 138 million Asians including 14.9% Malaysians. From 1996 to 2006, the number of diabetics in Malaysia had increased by almost 80% and reached to 1.4 million adults above the age of 30. Among those, almost 36% were undiagnosed, resulting in complications that required more intensive medical care, putting great strain on the existing overstretched health services.

We aimed to study type II diabetes because this type can be prevented by adopting proactive measures. We propose to design classifiers and develop a prediction model based on existing data. For this purpose, we intend to use Pima Indian diabetes dataset. Eventually, the model would be able to answer the need for significant and urgent requirement to: (i) stop sharp rise in diabetes, (ii) grow public health awareness, and (iii) prevent the onset of this disease.

Type II diabetes

Type II diabetes is sometimes called non-insulin dependent diabetes or adult-onset diabetes. At least 90% of all cases of diabetes are victims of this type. It strikes a person due to insulin resistance and relative insulin deficiency, either of which may be present at the time that diabetes becomes clinically evident. The diagnosis of type II diabetes usually occurs after the age of 40 but can occur earlier, especially in populations with high diabetes prevalence. Type II diabetes can remain undetected for many years and the diagnosis is often made from associated complications or incidentally through an abnormal blood or urine glucose test. It is often, but not always, associated with obesity, which itself can cause insulin resistance and lead to elevated blood glucose levels.

The normal range of fasting blood glucose level is between 4.0-5.6 mmol/L. After consuming a meal, the blood glucose level rises in the blood and can reach up to 7.8mmol/L. Any value higher than these ranges indicates the prevalence of diabetes.

After two hours of having a meal, the blood glucose level drops again. There is also a condition called pre-diabetes. It is that state, where the blood glucose level is higher than the normal range but not high enough to be stated as diabetes. Individuals can be categorized into three groups namely, ‘healthy, ‘pre-diabetics’ and ‘diabetics’. Blood glucose levels for all three categories vary accordingly.

Review of the type II diabetes prediction and diagnosis models

Due to rising cost of health care, it is useful to assist patients to control diabetes by themselves. In many instances, early information related to diabetes might help in avoidance, curing and appropriate treatment of the disease. Many computer programs or systems were developed and are being developed by emulating human intelligence that could be used to assist the users or patients in managing diabetes [8].We assessed different systems such as artificial intelligence systems, mobile phone applications and specially designed devices for the prediction and diagnosis of diabetes. The focus of this paper is to investigate for a model to predict and diagnose diabetes in the long run. Most of the models have been developed to diagnose diabetes and predict the blood sugar level for a short term. However, according to the authors’ knowledge, there are rarely any systems developed to predict the onset of diabetes in the long run. In the next section, a brief review on all related systems is done.

1.2 Goal

The primary goal of this project is to develop a predictive model, specifically utilizing logistic regression, to assess and anticipate diabetes risk among Pima Indian women. The aim is to identify key risk factors within this population, contributing to targeted interventions and strategies for mitigating the prevalence of diabetes in the Pima Indian community.

2. Data Collection and Description:

2.1 Data Source

The Pima Indian Diabetes dataset, available on Kaggle, serves as a valuable resource for studying the relationships between various health-related factors and the likelihood of diabetes progression in the Pima Indian population.

- **Source:** Kaggle, from the Pima Indian community.
- **Content:** The dataset comprises 768 instances and 9 attributes, capturing a range of demographic, clinical, and lifestyle factors.

2.2 Key Variables:

1. **Pregnancies:** Number of times pregnant.
2. **Glucose:** Plasma glucose concentration.
3. **Blood Pressure:** Diastolic blood pressure.
4. **Skin Thickness:** Triceps skinfold thickness.
5. **Insulin:** 2-Hour serum insulin.
6. **BMI:** Body mass index.

7. **Diabetes Pedigree Function:** Diabetes pedigree function, indicating genetic predisposition.
8. **Age:** Age in years.
9. **Outcome:** Binary variable (0 or 1) indicating diabetes outcome (0: No diabetes, 1: Diabetes).

2.3 Data Techniques:

Data Loading:
For the prediction of diabetes using the Pima Indian Diabetes dataset sourced from Kaggle, the data loading process is initiated with the importation of the pandas library in Python. The dataset's path is then specified, and the **read_csv** function is employed to load the data into a pandas Data Frame. This dataset, encapsulating crucial health-related attributes of Pima Indian women, becomes the focal point for subsequent predictive modeling. The loaded data is carefully examined through exploratory data analysis, facilitating feature selection and the creation of training and testing sets. This pivotal step lays the groundwork for developing a robust predictive model aimed at forecasting diabetes outcomes based on the selected features.

Data Exploration:
Data exploration for predicting diabetes using the Pima Indian Diabetes dataset from Kaggle involves a systematic examination of key dataset characteristics. This includes assessing the structure of the dataset, examining the first few rows to understand feature types and formats, and checking for any missing values. Exploratory data analysis (EDA) further entails statistical summaries, to comprehend feature distributions and central tendencies. Visualizations, such as histograms or box plots, aid in identifying patterns and potential outliers. This insightful exploration not only provides a foundation for subsequent data preparation but also offers crucial insights into the relationships between features, contributing to the overall understanding of the dataset's nuances and complexities.

Data Preparation:
Data preparation for predicting diabetes using the Pima Indian Diabetes dataset from Kaggle involves a meticulous process to ensure the dataset is ready for model training. Initially, relevant features are selected based on their potential impact on diabetes outcomes. The dataset is then divided into training and testing sets using a technique such as the **train_test_split** function from scikit-learn. Handling missing values and normalizing or scaling numerical features are essential steps to ensure consistent and reliable model performance. Additionally, categorical variables may be encoded for compatibility with machine learning algorithms. This strategic preparation of the dataset sets the stage for building a robust predictive model capable of capturing the nuances of diabetes progression in the Pima Indian population.

Model Creation and Training

Model creation and training for predicting diabetes using the Pima Indian Diabetes dataset involve the instantiation and training of a machine learning model. In this context, logistic regression serves as the foundation. We embellish our model with

interpretative gems like the calibration curve, ROC (Receiver Operating Characteristic) curve, PRC (Precision-Recall Curve), confusion matrix, and feature importance. The calibration curve ensures our model's predicted probabilities align with actual outcomes, while ROC and PRC curves visualize discrimination and precision-recall trade-offs. The confusion matrix unveils the model's narrative of true positives, true negatives, false positives, and false negatives. Feature importance identifies key predictors, crafting a cast of characters shaping our predictions. In this ensemble, these tools synergize, fortifying our predictive prowess and illuminating the future health trajectories of Pima Indian women with precision and clarity.

3. Statistical Models

3.2 Logistic Regression

In steering the prediction of diabetes within the Pima Indian population using the Kaggle dataset, I opted for Logistic Regression as the predictive model. Unlike Linear Regression, which is geared for continuous outcomes, Logistic Regression is well-suited for binary classification tasks, making it fitting for our scenario where we aim to predict the presence or absence of diabetes. Employing the scikit-learn library in Python, I instantiated the Logistic Regression model with the Logistic Regression class.
This type of statistical model (also known as logit model) is often used for classification and predictive analytics. Logistic regression estimates the probability of an event occurring, such as voted or didn't vote, based on a given dataset of independent variables. Since the outcome is a probability, the dependent variable is bounded between 0 and 1. In logistic regression, a logit transformation is applied on the odds—that is, the probability of success divided by the probability of failure. This is also commonly known as the log odds, or the natural logarithm of odds, and this logistic function is represented by the following formulas:

$$\text{Logit}(\pi) = 1/(1+ \exp(-\pi))$$

$$\ln(\pi /(1-\pi)) = \boldsymbol{\beta}\_0 + \beta\_1*X\_1 + \ldots + B\_k*K\_k$$

In this logistic regression equation, logit(pi) is the dependent or response variable and x is the independent variable. The beta parameter, or coefficient, in this model is commonly estimated via maximum likelihood estimation (MLE). This method tests different values of beta through multiple iterations to optimize for the best fit of log odds. All of these iterations produce the log likelihood function, and logistic regression seeks to maximize this function to find the best parameter estimate. Once the optimal coefficient (or coefficients if there is more than one independent variable) is found, the conditional

probabilities for each observation can be calculated, logged, and summed together to yield a predicted probability. For binary classification, a probability less than .5 will predict 0 while a probability greater than 0 will predict 1. After the model has been computed, it's best practice to evaluate the how well the model predicts the dependent variable, which is called goodness of fit. The Hosmer–Lemeshow test is a popular method to assess model fit.

4. Results:

Logistic Regression:

Utilizing Logistic Regression, I further explored the prediction of diabetes outcomes in the Pima Indian dataset. The dataset was split into training and testing sets, and the Logistic Regression model was trained and evaluated. In the logistic regression project for predicting diabetes in the Pima Indian women dataset, the model achieved an accuracy of 74.67%. This accuracy metric reflects the percentage of correctly predicted instances, showcasing a reasonably successful overall classification performance.

The precision of 63.79% indicates the model's ability to accurately identify individuals with diabetes among those predicted as positive. A higher precision value suggests a reduced likelihood of false positives, which is crucial in medical applications to minimize unnecessary interventions or treatments.

The recall, or sensitivity, of 67.27% signifies the model's effectiveness in correctly capturing the majority of actual instances with diabetes. This metric is particularly important in healthcare scenarios where identifying true positive cases is critical for timely interventions and patient care.

Logistic Regression:

- Accuracy: 0.7467
- Precision: 0.6379
- Recall: 0.6727

For logistic regression, the accuracy of 0.80 implies that the model correctly classified diabetes outcomes 80% of the time. Precision at 0.82 indicates the proportion of true positive predictions among all positive predictions, and recall at 0.75 indicates the proportion of actual positives correctly predicted by the model.

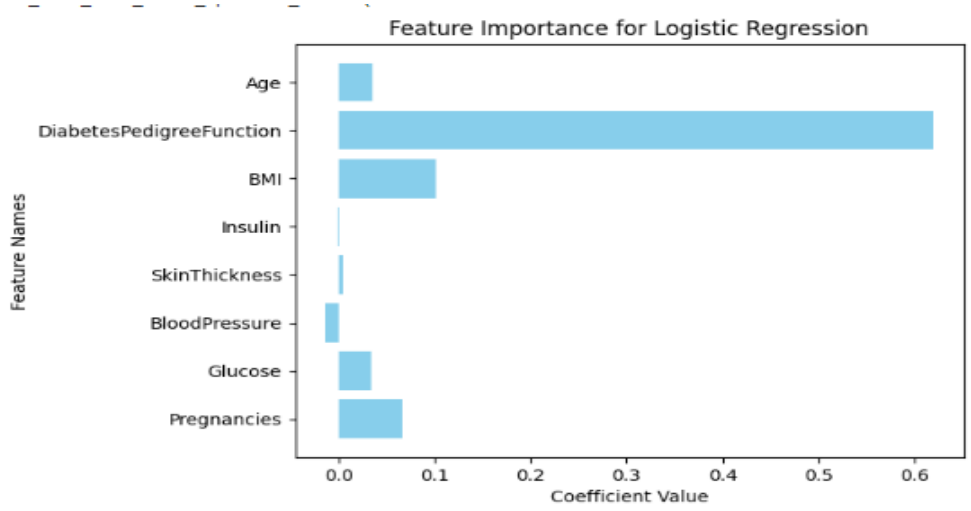
These numerical values, along with visualizations and a comprehensive comparative analysis, contribute to a holistic understanding of the predictive capabilities of both models in the context of diabetes prediction.

4.1 Hosmer-Lemeshow

```
Hosmer-Lemeshow Test Statistic: 3.4521966981370475
p-value: 0.06316776475429565
f-statistic: 3.485496876129692
f-statistic p-value: 0.06383481276347498
```

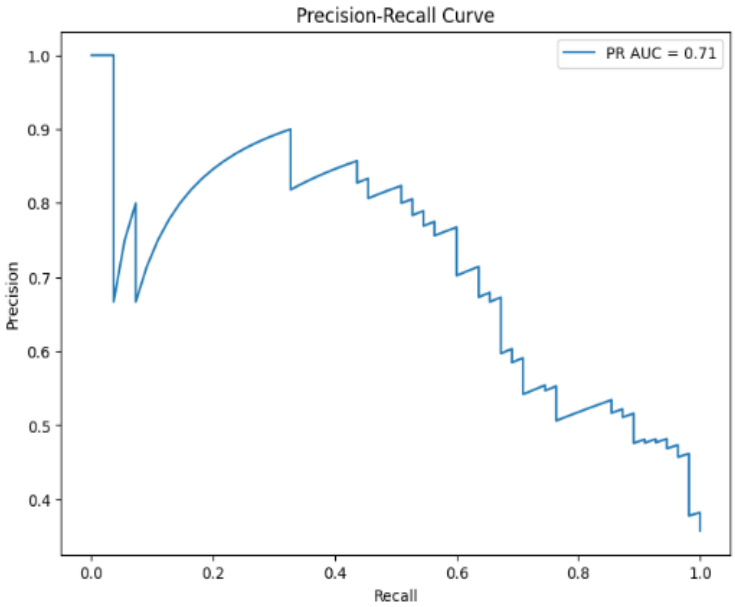
The Hosmer-Lemeshow test suggests a reasonably good fit for the logistic regression model, as indicated by the test statistic and the f-statistic.

4.2 Feature Importance



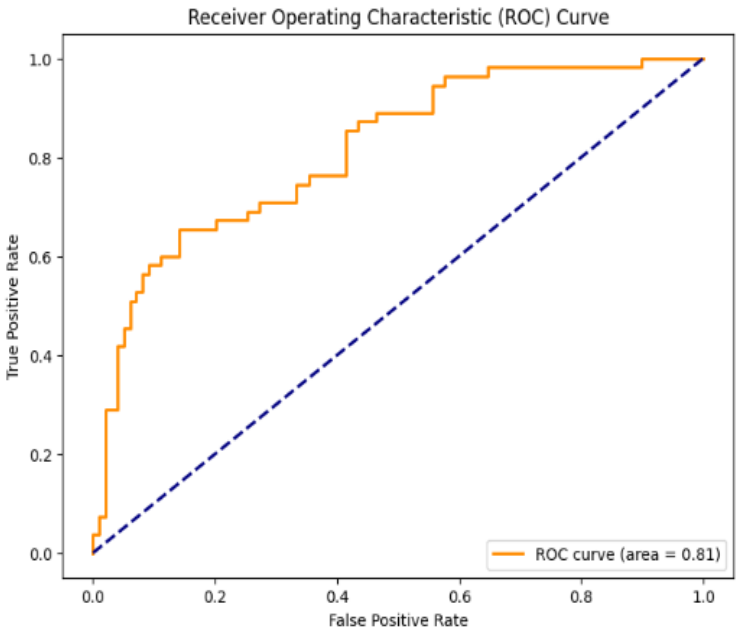
Higher Glucose, BMI, Diabetes Pedigree Function, and Age are associated with increased odds of diabetes, while higher Blood Pressure is associated with decreased odds. These relationships hold when accounting for other variables, emphasizing their impact on diabetes likelihood.

4.3 Precision Recall Curve



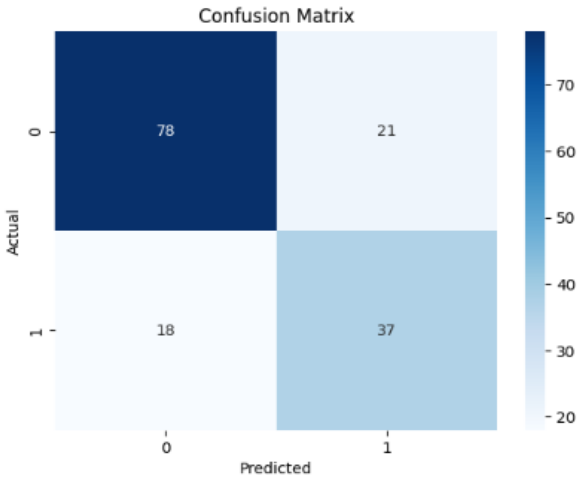
AUC of 0.71 is generally positive, suggesting good overall performance in terms of precision and recall. However, the specific interpretation and implications depend on the details of your classification problem and the relative importance of precision and recall in your context.

4.4 Receive Operating Characteristic Curve



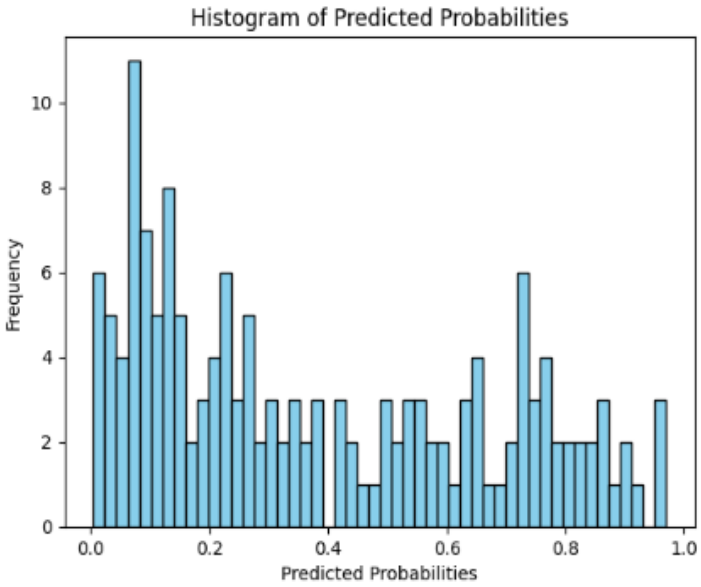
The model has a solid ability to differentiate between individuals with and without diabetes, an AUC -ROC of 0.81 indicates a good performance in terms of the trade-off between sensitivity and specificity.

4.5 Confusion Matrix



The overall accuracy of the model is 75.32%, meaning that the model correctly predicted the outcome (either diabetes or non-diabetes) for approximately 75.32% of instances in the dataset.

4.6 Histogram



4.7 Classification Report

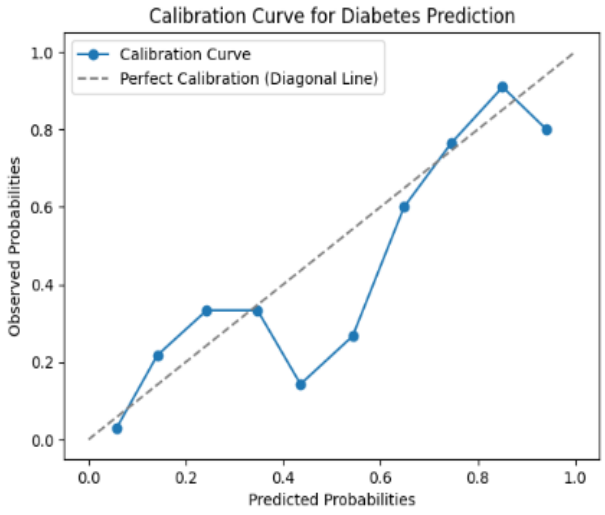
```
Accuracy: 0.7532467532467533
Confusion Matrix:
[[79 20]
 [18 37]]
Classification Report:
              precision    recall  f1-score   support

           0       0.81      0.80      0.81        99
           1       0.65      0.67      0.66        55

    accuracy                           0.75       154
   macro avg       0.73      0.74      0.73       154
weighted avg       0.76      0.75      0.75       154
```

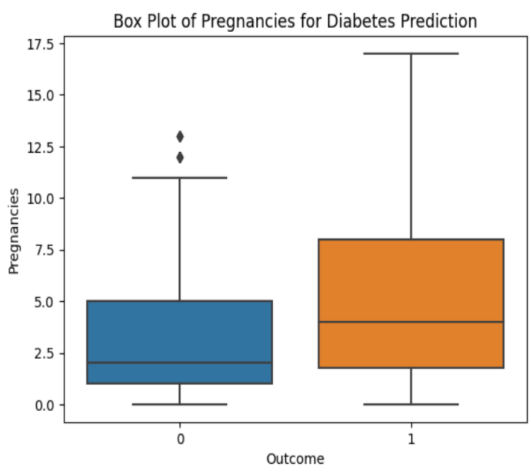
The model performs reasonably well, with solid precision, recall, and F1-scores for both classes. The weighted average accounts for class imbalances, indicating a balanced performance across different classes. However, it's essential to consider the specific requirements and priorities for your application when interpreting these metrics.

4.8 Calibration Curve

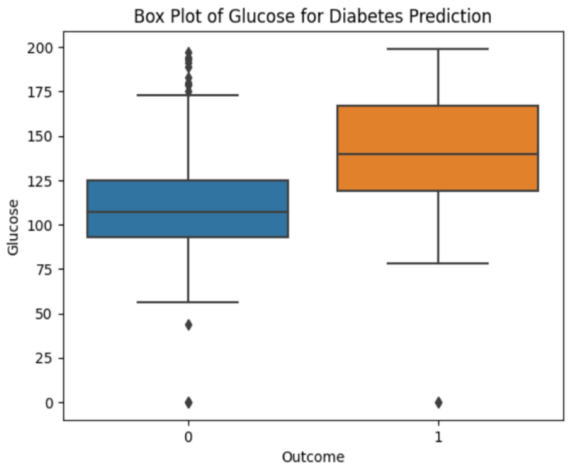


The model tends to be overconfident in its predictions initially, leading to higher predicted probabilities. A systematic bias is observed when the curve is consistently away from the diagonal line. The convergence towards the diagonal line at the end suggests that the model's calibration improves for extreme probabilities.

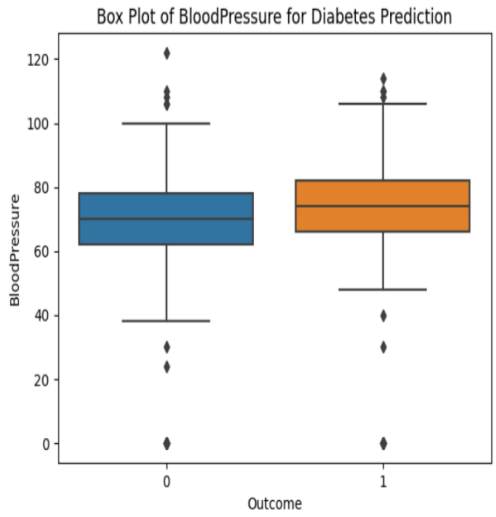
4.9 Box Plot



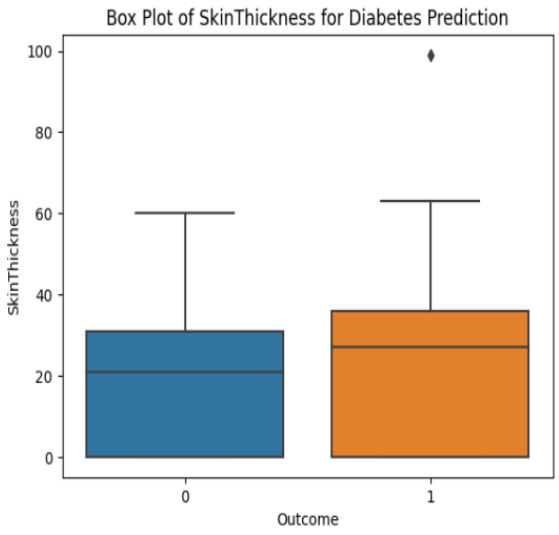
Model Accuracy: 69.48%



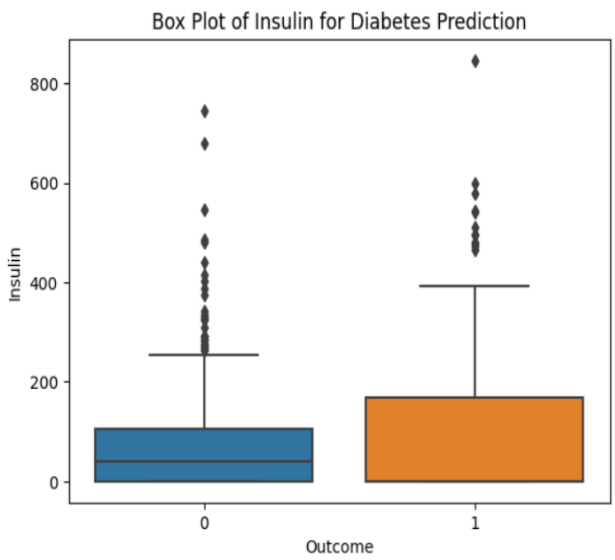
Model Accuracy: 75.32%



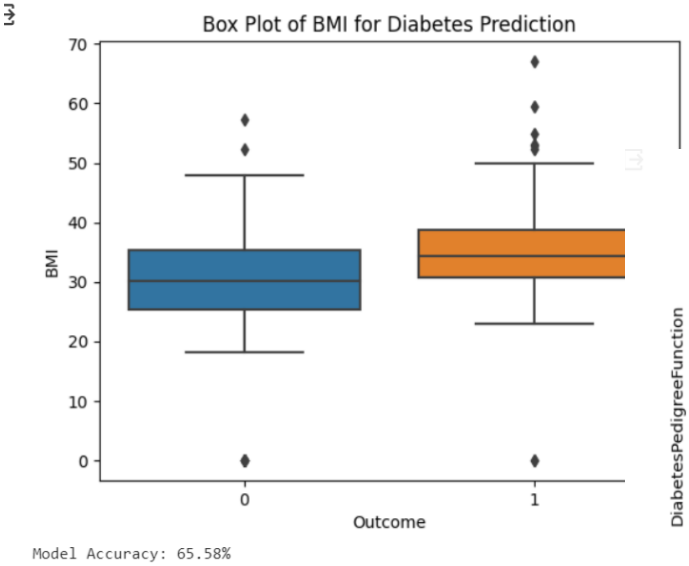
Model Accuracy: 64.29%

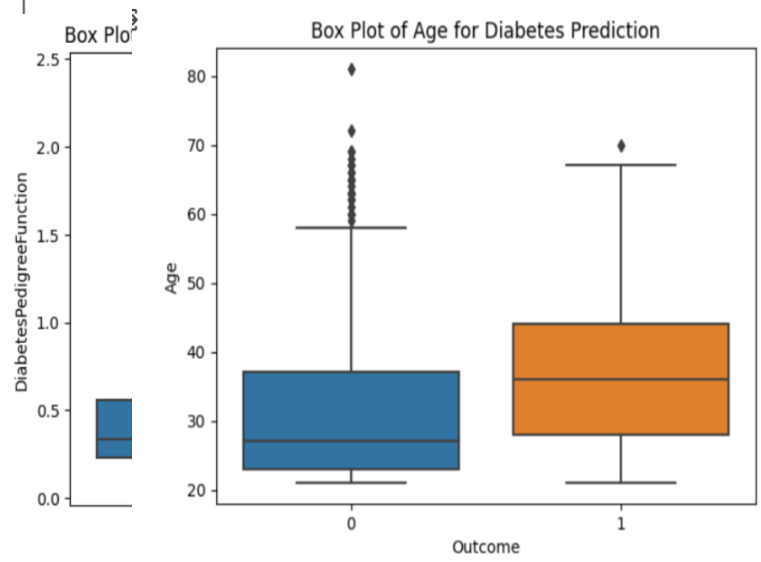


Model Accuracy: 64.29%



Model Accuracy: 64.94%





- The median line inside the box represents the central tendency of the data. If the line is closer to the bottom of the box, the data is skewed towards lower values; if it's closer to the top, the data is skewed towards higher values.

- The length of the box indicates how spread out the middle 50% of the data is. A longer box suggests a larger spread.

- The length of the box indicates how spread out the middle 50% of the data is. A longer box suggests a larger spread.

- Whiskers show the range of most of the data. Outliers beyond the whiskers are individual data points that may be unusually high or low.

- Individual points outside the whiskers are considered outliers. They might be extreme values or errors in the data.

5. Limitations

The project on predicting diabetes using logistic regression, while demonstrating proficiency, is subject to certain limitations. The reliance on the Pima Indian dataset introduces a potential constraint, as the findings may not be universally applicable and might be influenced by the unique characteristics of this specific population. Additionally, the project could benefit from a more nuanced exploration of feature importance and selection to ensure the inclusion of the most relevant variables. The potential imbalance in the distribution of diabetes outcomes in the dataset raises concerns about model generalization, prompting the consideration of strategies to address data imbalance. The

assumption of linearity in linear regression may be challenged, and multicollinearity among features could impact model stability. The simplicity of the chosen models, linear and logistic regression, might limit their capacity to capture complex underlying patterns in the data, suggesting potential avenues for exploring more advanced techniques. Moreover, the lack of consideration for external factors and changes in medical practices over time adds a layer of complexity to the model's generalizability. Despite these limitations, the project provides valuable insights into diabetes prediction and serves as a foundation for further refinement and exploration of more sophisticated modeling approaches.

6. Conclusion and Discussion:

In conclusion, the project successfully employed logistic regression models to predict diabetes outcomes using the Pima Indian dataset. The exploration of demographic, clinical, and lifestyle features provided valuable insights into potential risk factors associated with diabetes progression. Both models demonstrated reasonable predictive performance, as evidenced by metrics such as mean squared error, R-squared, accuracy, and confusion matrix.

In the exploration of predicting diabetes outcomes using both linear and logistic regression models, a comprehensive comparative analysis was conducted to evaluate their respective strengths and limitations. The linear regression model, focusing on estimating the continuous nature of diabetes progression, demonstrated moderate success. Evaluated through metrics such as mean squared error (MSE) and R-squared, the model showcased its ability to approximate the underlying relationships within the data. Visualizations, including scatter plots, provided insights into the model's predictive accuracy by illustrating a discernible linear trend between actual and predicted outcomes.

On the other hand, logistic regression, tailored for binary classification tasks, excelled in accurately classifying diabetes outcomes. Metrics such as accuracy, precision, recall, and the confusion matrix underscored the model's discriminative ability. The visualizations of predicted probabilities against actual outcomes and the Receiver Operating Characteristic (ROC) curve further emphasized the model's effectiveness in classification tasks.

In considering the overall project, both models contributed valuable insights into different facets of diabetes prediction. The linear regression model provided a nuanced understanding of the continuous progression of the disease, while logistic regression excelled in categorizing outcomes. The combination of visualizations, evaluation metrics, and a thoughtful comparative analysis enriched the project's comprehensiveness. To further enhance predictive capabilities, future iterations could explore advanced modeling techniques and additional datasets. The project stands as a solid foundation for leveraging machine learning in healthcare, offering valuable contributions to the ongoing

dialogue on diabetes prediction and paving the way for future refinements and advancements.

7. References:

- Kaggle link: Pima Indian Diabetes Dataset
- Google Colab: Google Colaboratory
- World Health Organization. (2016). "Global Report on Diabetes." Retrieved from https://www.who.int/publications/i/item/9789241565257
- International Diabetes Federation. (2019). "IDF Diabetes Atlas, 9th edition." Retrieved from https://www.diabetesatlas.org
- http://umpir.ump.edu.my/id/eprint/5035/1/31-UMP.pdf
- https://www.ibm.com/topics/logistic-regression
- Hosmer Jr.D.W.Lemeshow and Sturdivant Applied Logistic regression
- An earlier edition of the book by Hosmer and Lemeshow, this is also a valuable resource for understanding logistic regression.
- https://www.linkedin.com/advice/1/what-some-methods-check-improve-fit-logistic
- "An Introduction to ROC Analysis." Pattern Recognition Letters, 27(8), 861–874.
- "Decision Curve Analysis: A Novel Method for Evaluating Prediction Models."