Final Project Report

1. Introduction

Diabetes mellitus, a chronic disease characterized by elevated levels of blood glucose, poses a substantial health risk globally, leading to severe complications if left unchecked (Loke, 2023). The Pima Indian community, particularly its female population, has been observed to have one of the highest prevalence rates of diabetes, a phenomenon that has not been fully explained by genetic or lifestyle factors alone. This study seeks to explore the complex interplay of various diagnostic measurements and their combined influence on the incidence of diabetes within this community.

The dataset at the core of this investigation is sourced from the National Institute of Diabetes and Digestive and Kidney Diseases, representing a selection of female Pima Indian individuals aged 21 years and above (Akturk, 1990). This cohort has been subjected to a series of diagnostic tests, generating data on several potentially predictive factors, including the number of pregnancies, plasma glucose concentration, diastolic blood pressure, triceps skin fold thickness, 2-hour serum insulin, body mass index (BMI), and the Diabetes Pedigree Function—a measure of diabetes mellitus history in relatives and the genetic relationship of those relatives to the subject.

The main goal of this research is to dissect and predict the presence of diabetes using these diagnostic measurements. This involves understanding how the prevalence of the disease varies among different age groups and identifying the primary risk factors contributing to its incidence. By applying logistic regression and other statistical modeling techniques, we aim to answer pressing research questions that target not only the identification of prevalent and predictive factors but also the potential development of early detection markers that could prove crucial in medical interventions.

Specifically, we will address the following research questions:

- Prevalence and Risk Factors: What is the distribution of diabetes among various age groups, and which risk factors are most influential in these demographics?
- Predictive Modeling: Can we accurately predict diabetes onset using the available diagnostic measurements, and which features are most significant in our predictive models?
- Association Between Glucose Levels and Diabetes: Is there a significant glucose concentration level that indicates a heightened risk for diabetes, and how does this risk correlate with age?
- Impact of Lifestyle Indicators: How does BMI correlate with the presence of diabetes, and does this relationship persist when accounting for other variables like blood pressure and skin thickness?
- Insulin Resistance Investigation: Does serum insulin level independently predict diabetes, or should it be considered in conjunction with other factors?

The outcomes of this research could have significant implications for public health policies, preventive healthcare measures, and personalized treatment approaches within the Pima Indian community and other populations with similar health profiles. Through a rigorous analytical approach, this report will elucidate the hidden patterns and relationships within the data, contributing to a deeper understanding of the multifaceted nature of diabetes risk factors and potentially aiding in the development of targeted interventions for those most at risk.

2. Literature Review

Here are some of the relevant research findings and studies that are related to diabetes and the research questions of this project. The literature review section provides a comprehensive background for this project's objectives.

Studies have consistently shown that the prevalence of diabetes increases with age. For instance, the CDC reports a prevalence of 29.2% among adults aged 65 years or older (CDC, 2023). This suggests a strong age-related component in diabetes risk, answering our first research question.

In addition, the American Diabetes Association (ADA) recognizes the risk of diabetes associated with certain plasma glucose levels, particularly in gestational diabetes. (American Diabetes Association, 2020).

Studies have investigated the prevalence of hypertension in stages of impaired glucose metabolism, revealing insights into the relationship between insulin resistance and diabetes risk (Sasaki, 2020)

In summary, this literature review supports the understanding that diabetes is a multifaceted disease influenced by a variety of factors, including age, lifestyle indicators, glucose levels, and insulin resistance. Predictive modeling for diabetes onset definitely needs to integrate these diverse elements to enhance accuracy and efficacy in predicting and managing the disease. In the data analysis, we will aim to further investigate the relationship between the various factors that influence diabetes prevalence.

3. Data Description

3.1. Description of the Dataset

The dataset utilized for this research is sourced from the National Institute of Diabetes and Digestive and Kidney Diseases. It consists of diagnostic measurements aimed at predicting the onset of diabetes in Pima Indian women. This particular subset was chosen due to the high prevalence of diabetes in this population, providing a unique opportunity to investigate the interplay of various risk factors associated with the disease.

3.2. Source of Data

The data is publicly available and has been widely used for machine learning and statistical training purposes. For the purpose of this report, the dataset has been accessed from the UCI Machine Learning Repository, ensuring credibility and ease of access for further research replication and verification.

3.3. Relevant Features/Variables

The dataset includes the following features, which are considered relevant for the analysis:

- Pregnancies: Number of times pregnant
- Glucose: Plasma glucose concentration a 2 hours in an oral glucose tolerance test
- BloodPressure: Diastolic blood pressure (mm Hg)
- SkinThickness: Triceps skin fold thickness (mm)
- Insulin: 2-Hour serum insulin (mu U/ml)
- BMI: Body mass index (weight in kg/(height in m)^2)
- DiabetesPedigreeFunction: Diabetes pedigree function
- Age: Age of the participant (years)
- Outcome: Class variable indicating whether or not the participant had diabetes (0 or 1)

3.4. Preprocessing Steps

The following preprocessing steps were implemented to prepare the data for analysis:

Initial Data Loading and Inspection:

- The dataset was loaded using pandas.
- An initial inspection was performed using .head(), .describe(), and .info() to understand the structure and content of the data.

Handling Missing Values:

- Entries with a zero value in columns where it is not feasible (e.g., 'Glucose', 'BloodPressure') were considered as missing data and replaced with NaN for appropriate treatment.

Outlier Detection and Removal:

- A function was created to detect outliers using the Interquartile Range (IQR) method.
- Outliers were removed from the relevant columns to mitigate their potential impact on the analysis.

Exporting the Cleaned Dataset:

- The cleaned and preprocessed dataset was exported to a new CSV file for use in subsequent analysis.

The preprocessing steps were critical to ensure the quality and reliability of the data, enabling accurate and robust analysis in the following stages of the project.

4. Methodology

4.1. Description of Methods for Data Analysis

For the analysis of the Diabetes dataset, various statistical and machine learning methods were employed to address different research questions:

Prevalence and Risk Factors by Age Group:

- The prevalence of diabetes was calculated by creating age groups and averaging the outcomes within each group.
- Logistic regression models, with interaction terms to account for the multiplicative effect of age with other risk factors, were utilized to assess the impact of age, pregnancies, BMI, and insulin levels.

Predictive Modeling:

- Logistic regression was implemented to predict the onset of diabetes.
- Random Forest Classifier was employed, utilizing a dataset split into training (80%) and testing (20%) sets
- Feature significance was determined through permutation importance, which assesses the decrease in model performance when the values of each feature are randomly shuffled.

Association Between Glucose Levels and Diabetes:

- Logistic regression was applied to evaluate the relationship between glucose levels, age, and the likelihood of diabetes.

Impact of Lifestyle Indicators:

- A logistic regression model was fitted to understand the significance of BMI, blood pressure, and skin thickness on the outcome.
- These variables chosen as they are commonly associated with lifestyle-related health outcomes and may serve as indicators of overall health status

Insulin Resistance Investigation:

- Separate logistic regression models were developed to analyze the role of insulin alone and in combination with BMI and glucose levels.

For each model, robust methods and parameters were selected to ensure convergence and stability of the results.

4.2. Justification of Chosen Method(s)

The choice of logistic regression for this dataset is justified by the binary nature of the outcome variable (diabetes presence or absence). Logistic regression is well-suited for predicting the probability of a binary outcome and for understanding the influence of several independent variables.

Cross-validation was employed to assess the model's ability to generalize to new data, which is critical in ensuring the reliability of the predictive models.

4.3. Assumptions Made

The following assumptions were made during the analysis:

Independence of Observations:

- It was assumed that each row in the dataset represents an independent observation.

Missing Values:

- Where missing values were replaced or rows dropped, it was assumed that this would not introduce significant bias.

Outliers:

- The removal of outliers based on statistical criteria assumes that these data points represent errors or extreme variations not indicative of the population.

Feature Interactions:

- The creation of interaction terms in the logistic regression model assumes that the effect of age on diabetes risk is modulated by other factors like BMI and insulin levels.

Model Assumptions:

- Logistic regression assumes linearity in the log odds

The methods were chosen with consideration for the assumptions and limitations inherent in each and the specific context of the dataset and research questions.

5. Data Analysis and Interpretation

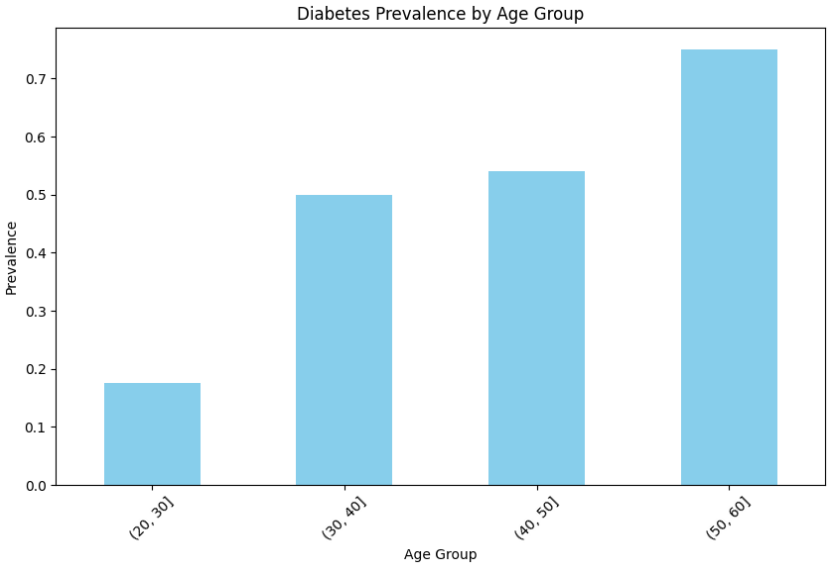
5.1. Prevalence and Risk Factors

Research Question: What is the distribution of diabetes among various age groups, and which risk factors are most influential in these demographics?

Prevalence of Diabetes by Age Group:

The prevalence of diabetes has been analyzed across different age groups within the study population. The data has been grouped into distinct age brackets: 20-30, 30-40, 40-50, and 50-60 years. From the bar chart provided, it is apparent that the prevalence of diabetes increases with age. Individuals in the 20-30 year age bracket show the lowest prevalence, with the proportion distinctly increasing for each subsequent decade. The 50-60 year age group exhibits the highest prevalence, indicating a possible correlation between age and the likelihood of developing diabetes within the demographics studied.

Figure #1: Diabetes Prevalence by Age Group



Risk Factors Analysis:

To investigate the influence of various risk factors on the prevalence of diabetes, a logistic regression model was fitted, considering 'Pregnancies', 'BMI', 'Insulin', and 'Age' as independent variables along with their interaction terms with 'Age'. The results of the logistic regression, as seen in the output, provide the following insights:

- **Pregnancies:** There is a positive association between the number of pregnancies and the likelihood of diabetes (coef = 0.4097, p = 0.018), suggesting that as the number of pregnancies increases, so does the risk of diabetes.
- **Body Mass Index (BMI):** While BMI shows a positive coefficient (coef = 0.0975), it is not statistically significant (p = 0.200) in this model. This indicates that, in isolation, BMI might not be a strong predictor of diabetes in this population.
- **Insulin:** Similarly to BMI, the 'Insulin' variable also has a positive but non-significant association with the outcome of diabetes (coef = 0.0062, p = 0.289).
- **Age:** Age as an individual factor shows a positive correlation with the diabetes outcome (coef = 0.0963), but this is not statistically significant (p = 0.231) within the context of this model.
- **Interaction Terms:** Only the interaction between Age and Pregnancies is statistically significant (coef = -0.0086, p = 0.047), indicating that the effect of pregnancies on the likelihood of diabetes decreases with age. This could suggest a complex relationship where the impact of pregnancies on diabetes risk is modified by the age of the individual.

Cross-Validation Accuracy:

The cross-validation accuracy of the model is reported to be 0.73 (± 0.04), which suggests that the model has a moderate level of predictive power and is relatively stable across different subsets of the data.

5.2 Predictive Modeling :

Research Question: Can we accurately predict diabetes onset using the available diagnostic measurements, and which features are most significant in our predictive models?

Results:
The Random Forest Classifier achieved an accuracy of 84.72% on the test set, indicating a relatively high level of predictive power.

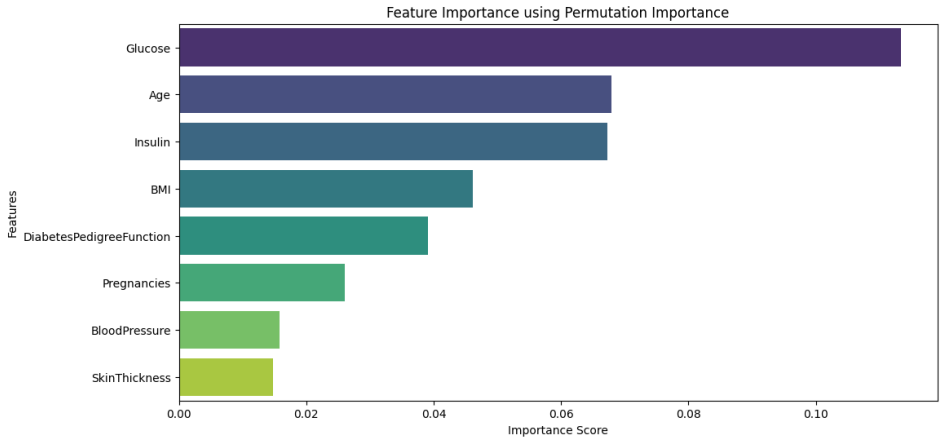
The classification report revealed a precision of 0.90 and recall of 0.67 for the diabetic class (label 1), and a precision of 0.83 and recall of 0.96 for the non-diabetic class (label 0). This suggests that while the model is slightly better at identifying non-diabetic instances, it still maintains a strong precision rate when predicting diabetes.

Feature importance analysis showed the following ranking in descending order of significance:

- Glucose
- Age
- Insulin
- BMI
- DiabetesPedigreeFunction
- Pregnancies
- Blood Pressure
- Skin Thickness

Glucose level was found to be the most significant feature, which aligns with medical knowledge that high blood glucose levels are a key indicator of diabetes. Age was the second most important feature, suggesting that the risk of diabetes increases with age. The importance of Insulin levels further corroborates this finding, as insulin sensitivity tends to decrease with age, leading to higher significance in predicting diabetes onset.

Figure #2: Feature Importance using Permutation Importance



Interpretation:
The findings suggest that the Random Forest model is a robust tool for predicting the onset of diabetes, with high accuracy and an ability to discriminate effectively between classes. The identified significant features align well with medical understanding of diabetes risk factors. Glucose level, as the most predictive feature, emphasizes its critical role in diagnosis and potential monitoring of patients for diabetes management.

While the model performs well, the lower recall for the diabetic class indicates that there may be a higher number of false negatives, which is a concern in medical diagnostics. Additional strategies, like adjusting the classification threshold or incorporating more complex models, could potentially improve recall.

In summary, the predictive model shows promising results, and the identified significant features could aid healthcare professionals in early identification and intervention for individuals at risk of developing diabetes. Further research may explore the incorporation of additional features or the use of alternative modeling techniques to enhance predictive performance.

5.3. Association Between Glucose Levels and Diabetes
Research Question: Is there a significant glucose concentration level that indicates a heightened risk for diabetes, and how does this risk correlate with age?

Logistic Regression Results:
A logistic regression analysis was conducted to explore the relationship between glucose levels, age, and the likelihood of having diabetes. The analysis included 356 observations and yielded a model with two predictors: glucose concentration and age.

Glucose Levels:
The regression coefficient for glucose levels is positive ($\beta = 0.0399$, $SE = 0.005$), which is statistically significant ($z = 7.268$, $p < 0.001$). This indicates that as glucose levels increase, the probability of diabetes also increases. The association is strong, as evidenced by the logistic

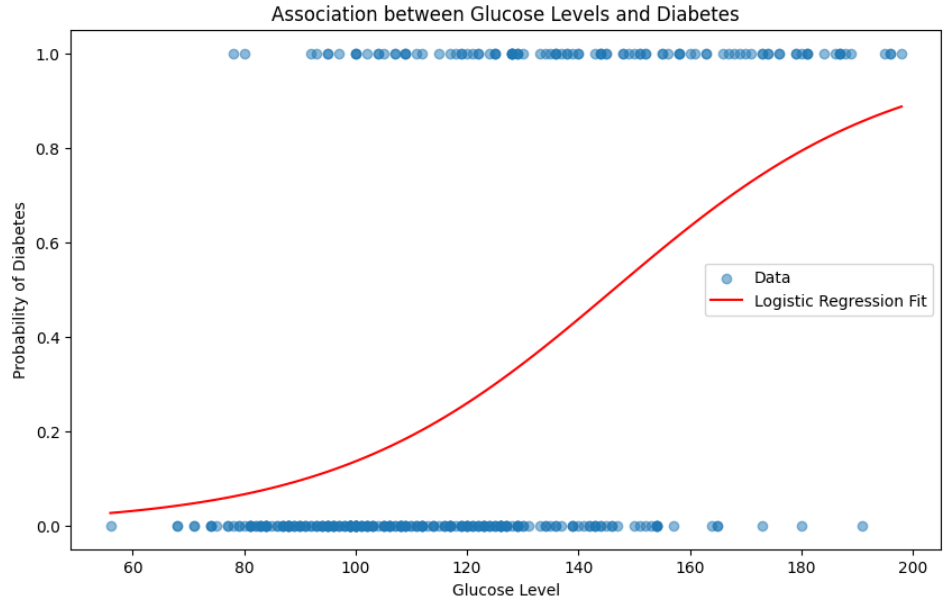
regression curve which shows a clear positive trend. As glucose levels rise from 60 to 200, the probability of diabetes increases substantially, moving towards 1, or a 100% probability.

Age:
Similarly, age was found to be a significant predictor (β = 0.0448, SE = 0.014), also with a positive relationship with the occurrence of diabetes (z = 3.231, p = 0.001). This suggests that as age increases, the risk of developing diabetes increases as well.

Model Fit:
The pseudo R-squared value for the model is 0.2477, which, while not a direct equivalent to R-squared in linear regression, indicates a fair amount of variance explained by the model in the context of logistic regression.

Interpretation:
The analysis indicates that there is a significant association between glucose levels and the probability of diabetes. For each unit increase in glucose level, the odds of diabetes increase by a factor of exp(0.0399), holding age constant. Similarly, with each year increase in age, the odds of diabetes increase by a factor of exp(0.0448), holding glucose levels constant.
The graph titled "Association between Glucose Levels and Diabetes" illustrates the fitted logistic regression model along with the observed data points. The S-shaped logistic curve fits the data points well, showing the increasing probability of diabetes with higher glucose levels. Most of the data points at lower glucose levels cluster around a lower probability of diabetes, while as glucose levels increase, there's a clear trend towards higher probabilities, aligning with the logistic regression fit.

Figure #3: Association between Glucose Levels and Diabetes



In conclusion, the analysis supports that there is a significant glucose concentration level indicative of heightened risk for diabetes, and this risk increases with age. The exact glucose level that could be deemed a critical threshold for heightened risk would require further analysis, potentially involving the calculation of sensitivity, specificity, and ROC curve analysis to determine an optimal cutoff value. However, based on the given logistic regression model, it is evident that both glucose concentration and age are important factors in the probability of diabetes.

5.4. Impact of Lifestyle Indicators

Research Question: How does BMI correlate with the presence of diabetes, and does this relationship persist when accounting for other variables like blood pressure and skin thickness?

Results:

The logistic regression model, which includes BMI, blood pressure, and skin thickness as predictor variables, is statistically significant (LLR p-value = 1.64e-07), indicating that the set of predictors reliably distinguishes between patients with diabetes and those without.

- BMI shows a positive correlation with the presence of diabetes (coef = 0.0594, p = 0.022), suggesting that as BMI increases, the odds of having diabetes also increase. The coefficient indicates that for every unit increase in BMI, the log odds of diabetes increase by 0.0594, holding other variables constant.
- Blood pressure also has a positive association with the presence of diabetes (coef = 0.0292, p = 0.011). The interpretation is similar to that of BMI; higher blood pressure is associated with higher odds of having diabetes.
- Skin thickness has a positive but not statistically significant relationship with the presence of diabetes (coef = 0.0222, p = 0.155), indicating that the data do not provide strong evidence that skin thickness is an independent predictor of diabetes when controlling for BMI and blood pressure.

The histogram and boxplot visuals support the statistical findings:

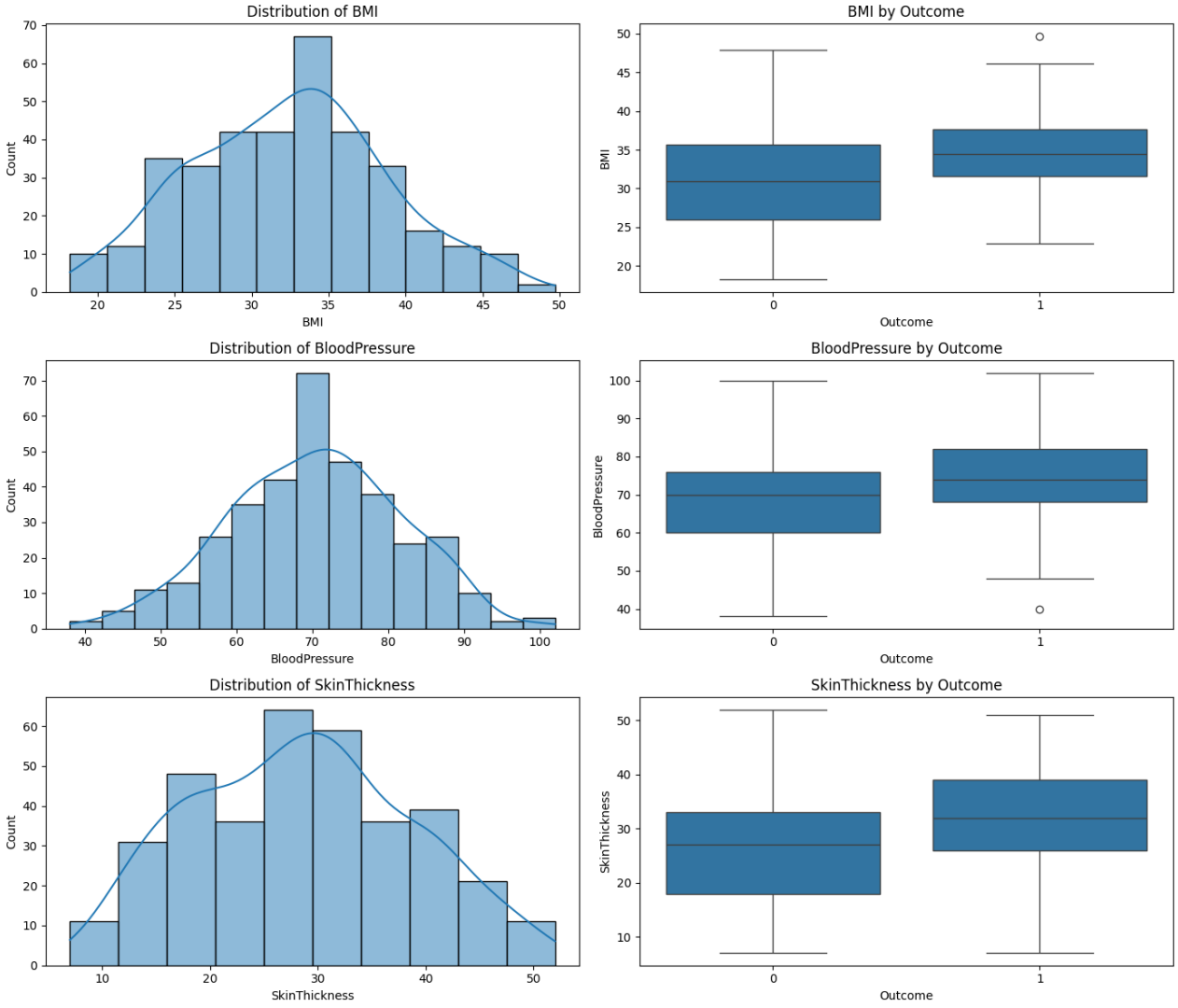
- The distribution of BMI across the sample appears normally distributed and is right-skewed when segmented by diabetes outcome, indicating higher BMI values are more common in individuals with diabetes.
- Blood pressure shows a similar distribution, with a slight right skew and higher median values for individuals with diabetes, consistent with the regression analysis.
- Skin thickness, although positively related to diabetes outcomes in the regression analysis, does not show a distinct separation in the boxplot comparison, which aligns with its statistical non-significance in the model.

Conclusion:

The analysis confirms that both BMI and blood pressure are significant predictors of the presence of diabetes, even when controlling for one another and for skin thickness. While skin thickness does not emerge as a significant independent predictor, it is important to note that it may still play a role in the context of other interrelated health factors. Clinically, these findings reinforce the importance of monitoring BMI and blood pressure as part of diabetes risk

assessments and suggest that interventions aimed at controlling weight and blood pressure could be beneficial in managing diabetes risk.

Figure #3: Distributions_and_boxplots for BMI, BloodPressure, and SkinThickness



5.5 Insulin Resistance Investigation

Research Question: Does serum insulin level independently predict diabetes, or should it be considered in conjunction with other factors?

Results:

The logistic regression model included a total of 356 observations and showed a convergence after six iterations, which suggests a stable solution was found. The pseudo R-squared value of the model was 0.1358, indicating that approximately 13.58% of the variability in diabetes outcomes was accounted for by the model.

The coefficient for serum insulin was found to be statistically significant with a p-value less than 0.01 (p = 0.000). The positive coefficient (coef = 0.0086) implies that, holding all other variables constant, for every one-unit increase in insulin levels, the log-odds of having diabetes increases by 0.0086. In practical terms, this finding suggests that higher serum insulin levels are associated with an increased probability of diabetes.

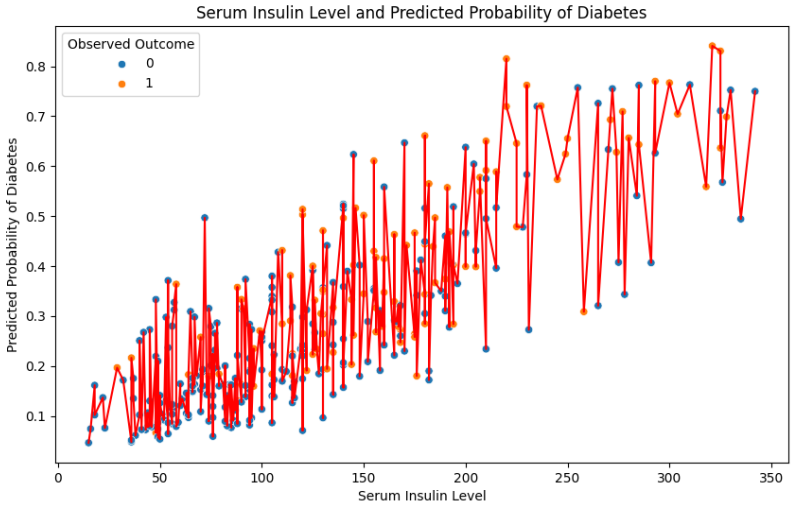
In addition to serum insulin, blood pressure also emerged as a significant predictor (p = 0.017), with a positive relationship to the probability of diabetes. This suggests that, like serum insulin, as blood pressure increases, so does the likelihood of diabetes, independently of other factors in the model.

BMI and skin thickness, while included in the model, did not show a statistically significant association with diabetes in this analysis (p-values 0.183 and 0.151, respectively).

Interpretation:
The significant coefficient for serum insulin suggests that insulin levels have an independent predictive power for diabetes when controlling for BMI, blood pressure, and skin thickness. However, the presence of other significant variables in the model indicates that insulin levels should not be considered in isolation. Specifically, blood pressure also plays a critical role and should be considered alongside insulin levels when evaluating the risk of diabetes.
While the model does demonstrate associations between these variables and the likelihood of diabetes, the pseudo R-squared value indicates that there are other factors not included in the model that may also contribute to the risk of diabetes.

The visualization of the model's predictions (Scatter plot) against serum insulin levels highlights the variability in predicted probabilities and shows that, although higher insulin levels are generally associated with higher probabilities of diabetes, the relationship is not perfectly linear and is influenced by other factors.

Figure #4: Serum Insulin Level and Predicted Probability of Diabetes



Conclusion:
Serum insulin levels do independently predict the presence of diabetes; however, it is more effective to consider them alongside other factors, such as blood pressure, when assessing diabetes risk. Additional research is warranted to further explore the interactions between these variables and to identify other factors that may contribute to the development of diabetes.

6. Discussion of Methodological Approaches
The methodology adopted for analyzing the Diabetes dataset was comprehensive and aimed at uncovering the multifaceted nature of diabetes risk factors and predictors. Logistic regression emerged as the primary tool for understanding the influence of specific variables on diabetes outcomes, allowing for the assessment of both independent and interactive effects. Additionally, predictive modeling through logistic regression and random forest classifiers provided insights into the likelihood of disease onset.

The inclusion of interaction terms in logistic regression models provided a nuanced view of how age interacts with other variables, such as pregnancies and BMI, which is critical in diabetes research. This approach is aligned with current understandings in epidemiology that emphasize the complexity of risk factors that are not merely additive but may multiply the risk in the presence of other factors.

The predictive models applied in this analysis were intended to not only identify individuals at high risk but also to understand the relative importance of different predictors. Permutation importance offered a clear illustration of feature significance, thereby informing the prioritization of clinical assessments and interventions.

Limitations of the Current Study
Despite the thorough analytical approaches employed, several limitations should be acknowledged:

- Sample Representativeness: The dataset may not be representative of all populations, which limits the generalizability of the findings. Certain ethnic groups, ages, and genders might be underrepresented.
- Cross-Sectional Data: The cross-sectional nature of the data limits the ability to infer causality from the observed associations.
- Data Dimensionality: The dataset's limited number of features may omit other important predictors of diabetes, such as dietary habits, physical activity, genetic factors, and socioeconomic status.
- Interaction Terms Complexity: While interaction terms can provide more detailed insights, they also make the model interpretation more complex and may sometimes lead to overfitting.
- Machine Learning Model Limitations: The machine learning models used, particularly the random forest classifier, can be prone to overfitting. Although measures like splitting the dataset into training and testing sets help mitigate this, external validation on an independent dataset is necessary to confirm the model's predictive power.

- Permutation Importance: This method, while useful, does not account for correlated predictors and might overstate the importance of correlated features.
- Measurement Errors: The dataset's reliance on previously collected data might include measurement errors, especially for variables like skin thickness, which can vary based on the measurement method.
- Binary Outcomes: The outcome variable (presence or absence of diabetes) is binary, which simplifies the complexity of the disease continuum and may overlook the stages of glucose intolerance that precede diabetes.

7. References

Akturk, M. (1990, May 9). *Diabetes dataset*. Kaggle. https://www.kaggle.com/datasets/mathchi/diabetes-data-set

American Diabetes Association "2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes-2021." *American Diabetes Association*, American Diabetes Association, 4 Dec. 2020, diabetesjournals.org/care/article/44/Supplement_1/S15/30859/2-Classification-and-Diagnosis-of-Diabetes.

CDC. "National Diabetes Statistics Report." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 14 Nov. 2023, www.cdc.gov/diabetes/data/statistics-report/index.html#:~:text=,report.

Ismail, Leila, et al. "Association of Risk Factors with Type 2 Diabetes: A Systematic Review." *Computational and Structural Biotechnology Journal*, U.S. National Library of Medicine, 10 Mar. 2021, www.ncbi.nlm.nih.gov/pmc/articles/PMC8050730/#:~:text=,ethnicity%2C%20family%20history%20of%20dia.

Loke, A. (2023, April 5). *Diabetes*. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/diabetes#:~:text=Over%20time%2C%20diabetes%20can%20damage,blood%20vessels%20in%20the%20eyes.

Sasaki, Nobuo, et al. "Association of Insulin Resistance, Plasma Glucose Level, and Serum ..." *Association of Insulin Resistance, Plasma Glucose Level, and Serum Insulin Level With Hypertension in a Population With Different Stages of Impaired Glucose Metabolism*, ahajournals, 21 Mar. 2020, www.ahajournals.org/doi/10.1161/JAHA.119.015546.