Initial Data Analysis:

Supervised Classification Algorithms for Early Detection of Diabetes

Team 3

|  |  |  |
| --- | --- | --- |
| **Alex Hunt**  School of Computer Science University of Oklahoma  DSA-5103-OL | **Oluchi Ejehu**  MPGE  University of Oklahoma  DSA-5103-OL | **Zainab Iyiola**  MPGE  University of Oklahoma  DSA-5103-OL |

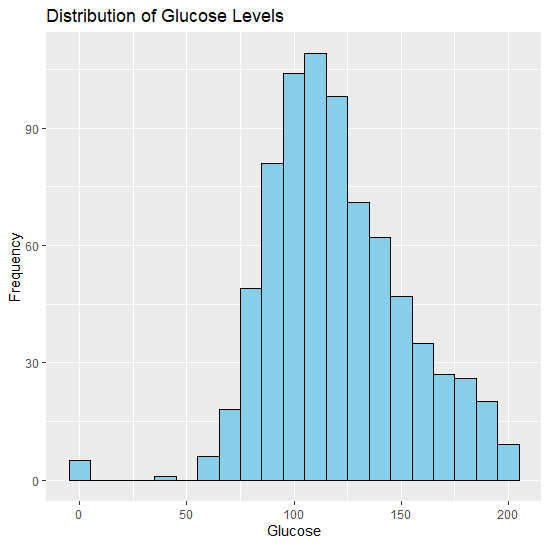
Professor Nicholson

November 6, 2024

**Approaches to Dealing with Data Issues**

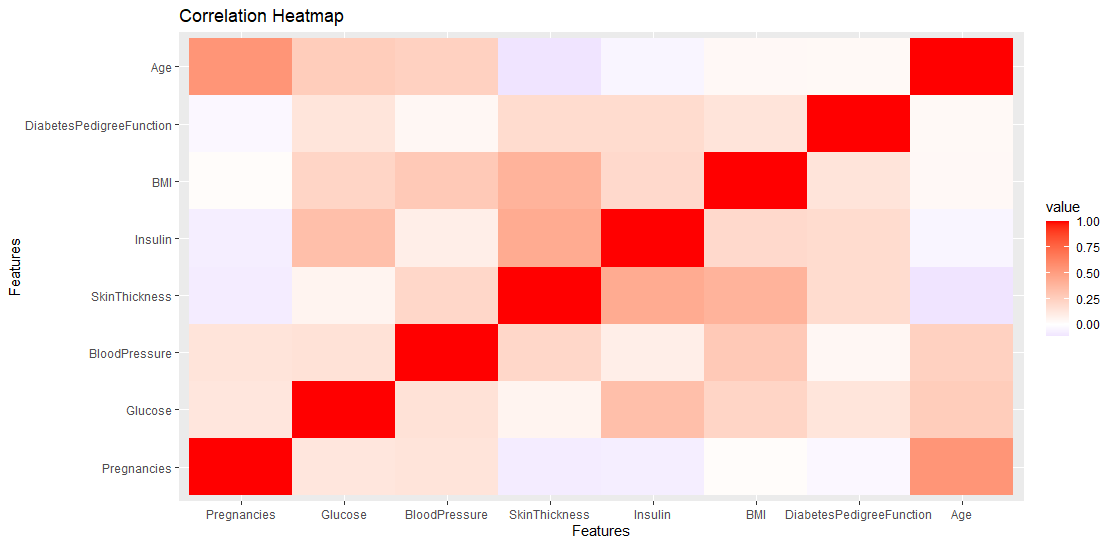
Our initial data analysis reveals several key data quality issues that we plan to address systematically to ensure the realism of our diabetes classification model. First, although our dataset does not contain missing values, we will remain vigilant about data imputation strategies should missing data arise in future iterations or extensions of the dataset. Outlier analysis, performed using the IQR method, identified several features with notable counts of outliers: BloodPressure (45), Insulin (34), and DiabetesPedigreeFunction (29). These outliers could potentially skew the model’s predictions, so we will consider strategies such as capping extreme values or applying transformations to mitigate their impact. Skewness is another critical issue, with features like BloodPressure, Insulin, DiabetesPedigreeFunction, and Age showing high skewness. We will explore transformation techniques, such as log or Box-Cox transformations, to normalize these distributions and improve model performance. Additionally, all features in our dataset are numeric, so we will not need to handle categorical factors. However, we will apply feature scaling to ensure consistent feature ranges and prevent model bias toward features with larger magnitudes. Overall, our approach focuses on preparing the data thoughtfully to maximize the effectiveness of our machine learning algorithms (see Table 1 in the Appendix).

**Insightful Visualizations**



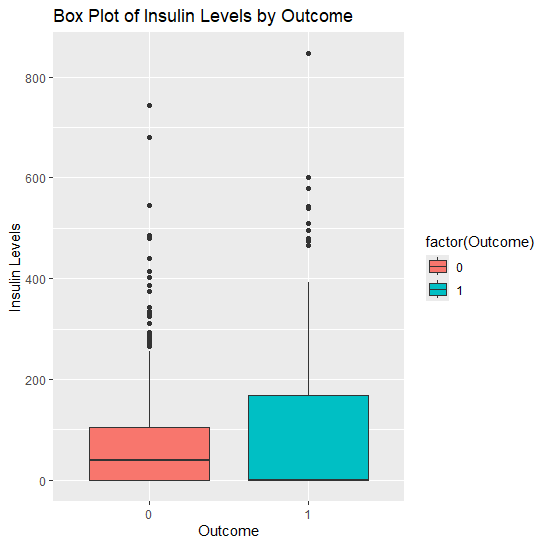
**Figure 1: Histogram of Glucose Levels**

The histogram of Glucose levels (Figure 1) was created using the ggplot2 library, with a bin width of 10 to display the distribution clearly. The histogram shows that the glucose values in the dataset are right-skewed, with a concentration of values between 70 and 150. The blue bars indicate the frequency of glucose measurements within each bin, helping us understand the spread and common ranges of glucose levels among patients. This visualization is crucial because glucose is a primary indicator used in diabetes diagnosis, and identifying its distribution helps us assess whether data transformations, like log normalization, are needed to improve model performance. Additionally, the histogram allows us to spot gaps or outliers that could impact the predictive power of our model.



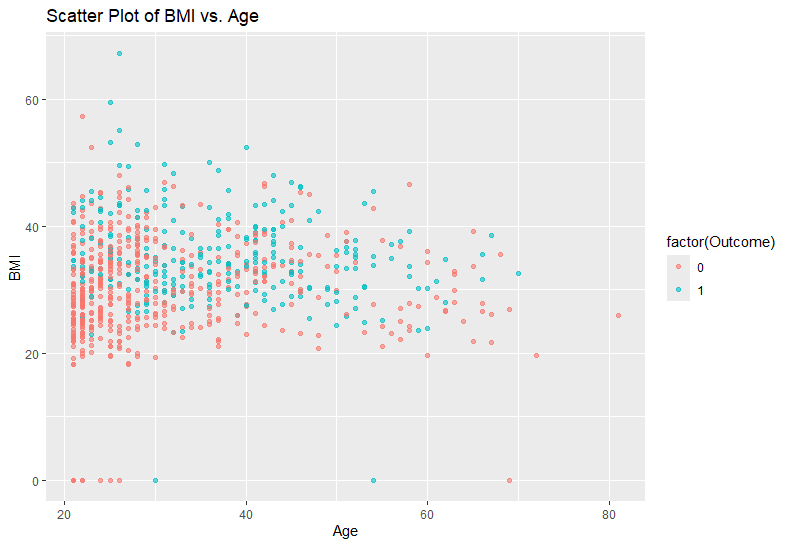
**Figure 2: Correlation Heatmap**

The correlation heatmap (Figure 2) was created by calculating the correlation matrix for all features (excluding Outcome) and then visualizing the matrix with the ggplot2 library. We used a gradient color scheme, where darker red indicates stronger positive correlations and darker blue shows stronger negative correlations, while white represents no correlation. This visualization is important because it helps us identify relationships between features that might affect our model's performance. For example, if two features are highly correlated, they could introduce multicollinearity, leading to less stable model estimates. By understanding these relationships, we can make informed decisions about feature selection, potentially removing or combining features to improve the model's predictive accuracy.



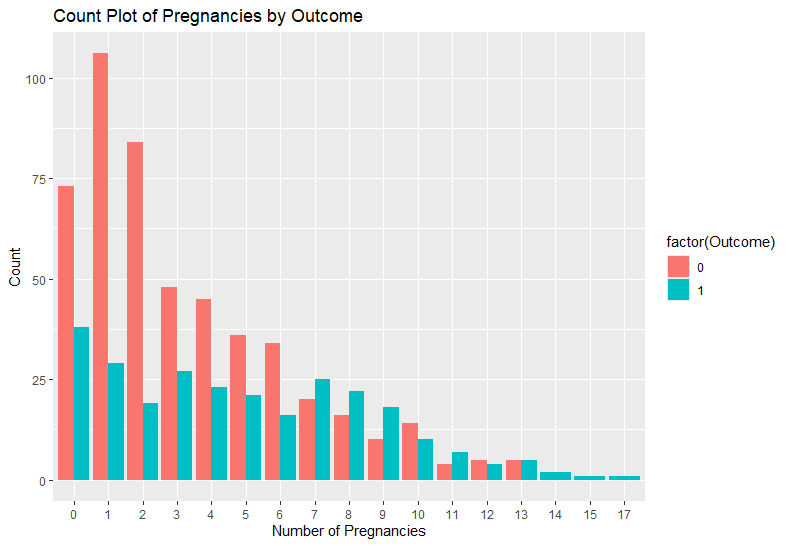
**Figure 3: Box Plot of Insulin Levels by Outcome**

The box plot of Insulin levels (Figure 3) was constructed using ggplot2, with Insulin on the y-axis and Outcome on the x-axis, separating the data into diabetic and non-diabetic groups. We also colored the boxes to visually differentiate the outcomes. The plot reveals that diabetic patients generally have higher insulin levels, though the distributions have significant variability and several outliers. This visualization is crucial because it highlights the role of insulin in diabetes prediction, emphasizing the need for careful outlier management. Addressing these outliers will be important for enhancing model robustness, as extreme values can skew model results and lead to inaccurate predictions.



**Figure 4: Scatter Plot of BMI vs. Age Colored by Outcome**

The scatter plot of BMI vs. Age (Figure 4) was created using ggplot2, with BMI on the y-axis and Age on the x-axis. The points were colored based on the Outcome variable, with diabetic and non-diabetic patients visually distinguished. We also used transparency (alpha) to prevent overplotting. This scatter plot reveals a trend where higher BMI values and older ages are often associated with diabetes. The importance of this visualization lies in its ability to show potential interactions between age and BMI, suggesting that both features might interact as risk factors. Recognizing these patterns will inform our modeling decisions, such as considering interaction terms or feature transformations to better capture these relationships.



**Figure 5: Count Plot of Pregnancies by Outcome**

The count plot of Pregnancies by Outcome (Figure 5) was generated using ggplot2, with the number of pregnancies on the x-axis and the counts separated by diabetes outcome. We used a dodge position to place bars side by side for a clearer comparison between diabetic and non-diabetic groups. The plot shows that women with higher numbers of pregnancies tend to have a greater likelihood of diabetes. This visualization is important because it emphasizes pregnancy frequency as a potential risk factor for diabetes in this population. Understanding this association helps us appreciate the medical relevance of the feature and informs our approach to modeling and interpreting the data.

**Appendix**

Table 1: Summary of Data Quality Issues

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Feature** | **Description** | **Missing Values** | **Outliers** | **Skewness** | **Factor** |
| Pregnancies | To express the Number of pregnancies | 0 | 4 | Moderate | No |
| Glucose | To express the Glucose level in blood | 0 | 5 | Low | No |
| BloodPressure | To express the Blood pressure measurement | 0 | 45 | High | No |
| SkinThickness | To express the thickness of the skin | 0 | 1 | Low | No |
| Insulin | To express the Insulin level in blood | 0 | 34 | High | No |
| BMI | To express the Body mass index | 0 | 19 | Low | No |
| DiabetesPedigreeFunction | To express the Diabetes percentage | 0 | 29 | High | No |
| Age | To express the age | 0 | 9 | High | No |
| Outcome | To express the final result 1 is Yes and 0 is No | 0 | 0 | Moderate | No |