Analysis of the Diabetes Dataset

March 20, 2025

1 Research Question

How do key clinical parameters—specifically plasma glucose concentration (Glucose), body mass index (BMI), and age—affect the likelihood of a positive diabetes test (Outcome) among Pima Indian women?

2 Data Analysis and Statistical Calculations

In our analysis, several descriptive and inferential statistical methods were used to understand the dataset and address the research question.

2.1 Descriptive Statistics

Feature	Count	Mean	Std Dev	Min	25%	50% (Median)	75%	Max
Pregnancies	768	3.85	3.37	0.0	1.00	3.00	6.00	17.00
Glucose	768	120.89	31.97	0.0	99.00	117.00	140.25	199.00
BP	768	69.11	19.36	0.0	62.00	72.00	80.00	122.00
ST	768	20.54	15.95	0.0	0.00	23.00	32.00	99.00
Insulin	768	79.80	115.24	0.0	0.00	30.50	127.25	846.00
BMI	768	31.99	7.88	0.0	27.30	32.00	36.60	67.10
DPF	768	0.47	0.33	0.078	0.24	0.37	0.63	2.42
Age	768	33.24	11.76	21.0	24.00	29.00	41.00	81.00
Outcome	768	0.35	0.48	0.0	0.00	0.00	1.00	1.00

Table 1: Descriptive Statistics of the Diabetes Dataset

For each variable in the dataset, the following statistics were calculated:

- **Central Tendency:** Mean, median (from the 50th percentile), and mode.
- **Dispersion:** Standard deviation and variance.

Feature	Count	Mean	Std Dev	Min	25%	50%	75%	Max	Mode
Pregnancies	768	3.85	3.37	0.0	1.00	3.00	6.00	17.00	1.00
Glucose	768	120.89	31.97	0.0	99.00	117.00	140.25	199.00	99.00
BP	768	69.11	19.36	0.0	62.00	72.00	80.00	122.00	70.00
ST	768	20.54	15.95	0.0	0.00	23.00	32.00	99.00	0.00
Insulin	768	79.80	115.24	0.0	0.00	30.50	127.25	846.00	0.00
BMI	768	31.99	7.88	0.0	27.30	32.00	36.60	67.10	32.00
DPF	768	0.47	0.33	0.078	0.24	0.37	0.63	2.42	0.254
Age	768	33.24	11.76	21.0	24.00	29.00	41.00	81.00	22.00
Outcome	768	0.35	0.48	0.0	0.00	0.00	1.00	1.00	0.00

Table 2: Descriptive Statistics of the Diabetes Dataset (Including Mode)

Feature	Skewness	Kurtosis
Pregnancies	0.9017	0.1592
Glucose	0.1738	0.6408
BloodPressure	-1.8436	5.1802
SkinThickness	0.1094	-0.5201
Insulin	2.2723	7.2143
BMI	-0.4290	3.2904
DiabetesPedigreeFunction	1.9199	5.5950
Age	1.1296	0.6432
Outcome	0.6350	-1.6009

Table 3: Skewness and Kurtosis of the Diabetes Dataset Features

- **Shape of the Distribution:** Skewness (to measure asymmetry) and kurtosis (to assess tail heaviness).

For example, the *Glucose* feature was observed to have a mean of approximately 120.9 with a standard deviation of about 32.0. The mode for Glucose was found to be 99, indicating a clustering of observations around this value.

2.2 Correlation Analysis

A Pearson correlation matrix was computed to examine the relationships among the features. Notable observations include:

- **Glucose and Outcome:** A correlation coefficient of roughly 0.47, indicating a moderate positive association. Higher glucose levels tend to be associated with a positive diabetes test.

Feature	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin
Pregnancies	1.0000	0.1295	0.1413	-0.0817	-0.0735
Glucose	0.1295	1.0000	0.1526	0.0573	0.3314
BloodPressure	0.1413	0.1526	1.0000	0.2074	0.0889
SkinThickness	-0.0817	0.0573	0.2074	1.0000	0.4368
Insulin	-0.0735	0.3314	0.0889	0.4368	1.0000

Table 4: Correlation Matrix (Part 1) - Pregnancies to Insulin

Feature	BMI	DPF	Age	Outcome
BMI	1.0000	0.1406	0.0362	0.2927
DPF	0.1406	1.0000	0.0336	0.1738
Age	0.0362	0.0336	1.0000	0.2384
Outcome	0.2927	0.1738	0.2384	1.0000

Table 5: Correlation Matrix (Part 2) - BMI to Outcome

- **BMI and Outcome:** A correlation of approximately 0.29, suggesting that higher BMI is also associated with a positive diabetes outcome.
- **Age and Outcome:** A correlation of about 0.24, implying that age has a positive, though slightly weaker, relationship with diabetes status.

2.3 Hypothesis Testing

To further explore the impact of the features on diabetes status, an independent two-sample t-test was performed comparing the Glucose levels between the two Outcome groups (0: negative, 1: positive). The test results were:

- **t-statistic:** -14.60
- **p-value:** < 0.0001

These results are highly statistically significant, indicating that the difference in mean Glucose levels between the groups is not due to random chance. The analysis confirms that patients who tested positive for diabetes have significantly higher Glucose levels.

3 Visualizations

A series of visualizations were used to further explore the dataset and illustrate the findings:

1. **Histograms:** Generated for each variable to inspect the characteristics of the distribution (including skewness, spread, and outliers). Figure: 1

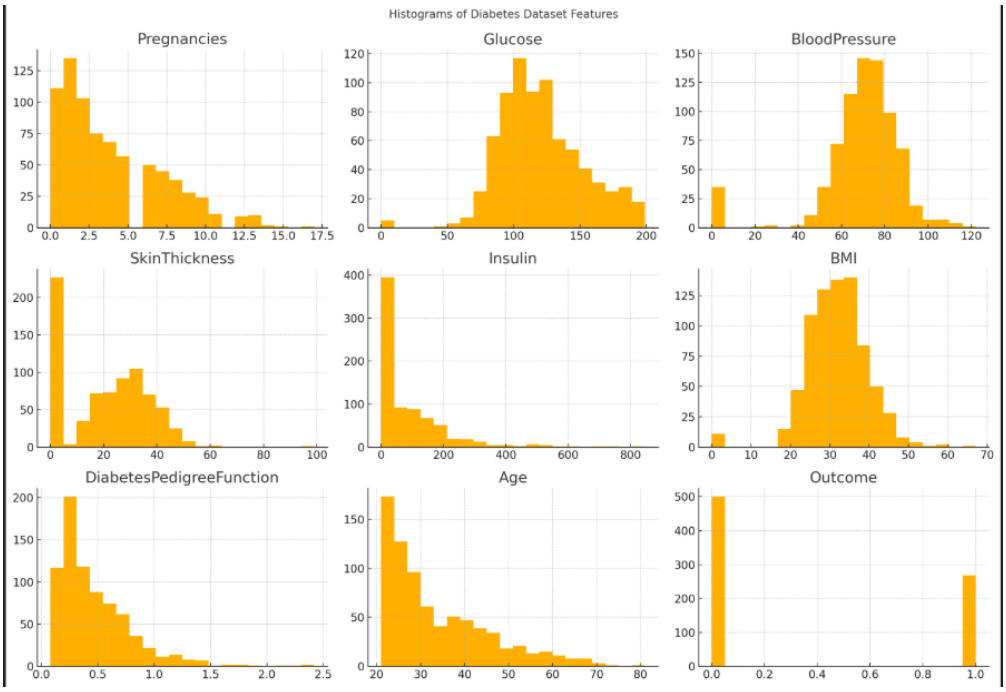

Figure 1: Histogram of Diabetes Dataset Features

2. **Boxplots by Outcome:** Boxplots of each feature (e.g., Glucose, BMI, Age) stratified by the Outcome variable. These plots highlighted differences in medians and spread between the groups.Figure: 2
3. **Scatterplot Matrix (Pairplot):** A pairplot was created using a subset of features (Glucose, BMI, Age, and Pregnancies) with Outcome as the hue. This visualization provided insight into pairwise relationships and potential clusters in the data. Figure: 3
4. **Correlation Heatmap:** A heatmap of the Pearson correlation matrix visually summarized the strength of the relationships among all variables. Figure: 4
5. **Density Plot:** A Kernel Density Estimate (KDE) plot for Glucose levels was used to illustrate the distribution differences between the Outcome groups. Figure: 5

4 Conclusion

The comprehensive analysis indicates that among Pima Indian women aged 21 and older:

- **Elevated plasma glucose levels** are strongly associated with a positive diabetes test.
- **Higher BMI and older age** are also associated with increased risk of diabetes, though the effect is most pronounced for Glucose.

The descriptive statistics, correlation analysis, and hypothesis testing collectively suggest that plasma glucose concentration is the most indicative of diabetes risk in this population. The visualizations further support these findings by illustrating the distinct differences in the distributions of key clinical parameters between the groups.

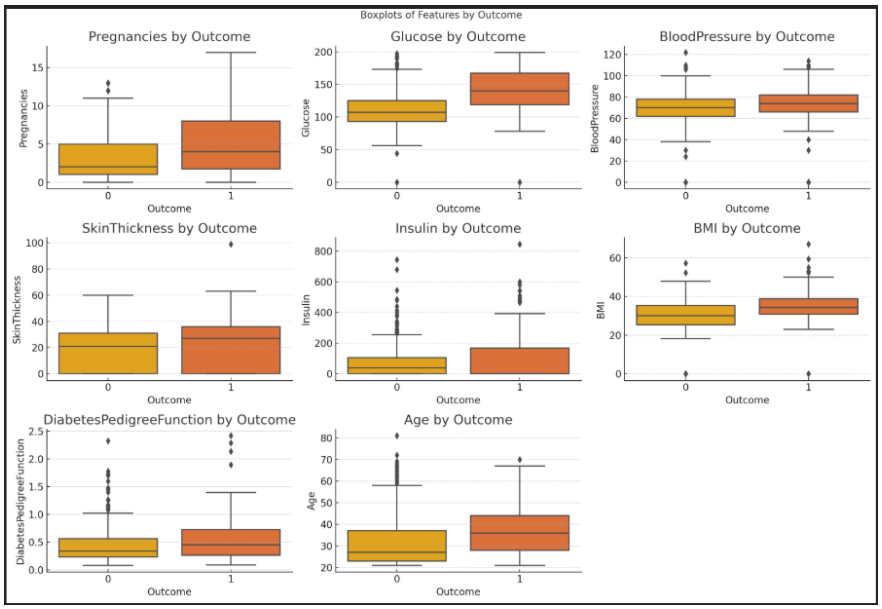

Figure 2: Boxplots of Features by Diabetes Outcome

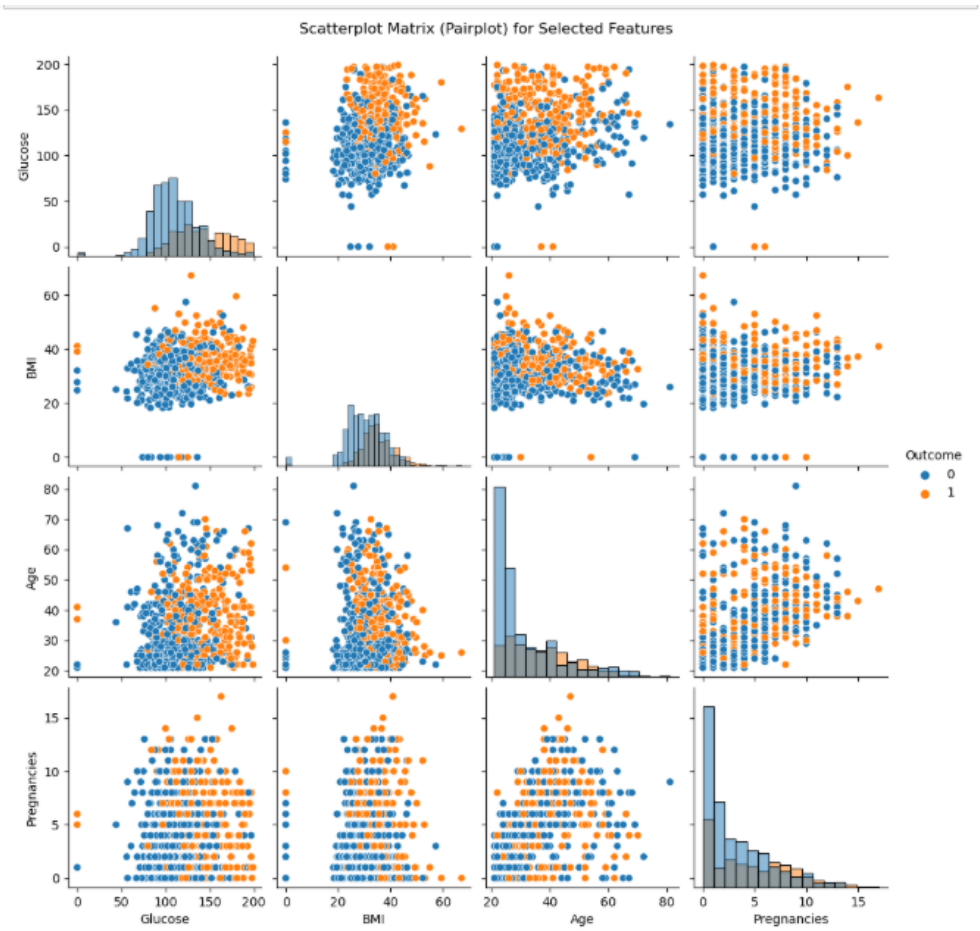

Figure 3: Scatterplot Matrix (Pairplot) for Selected Features

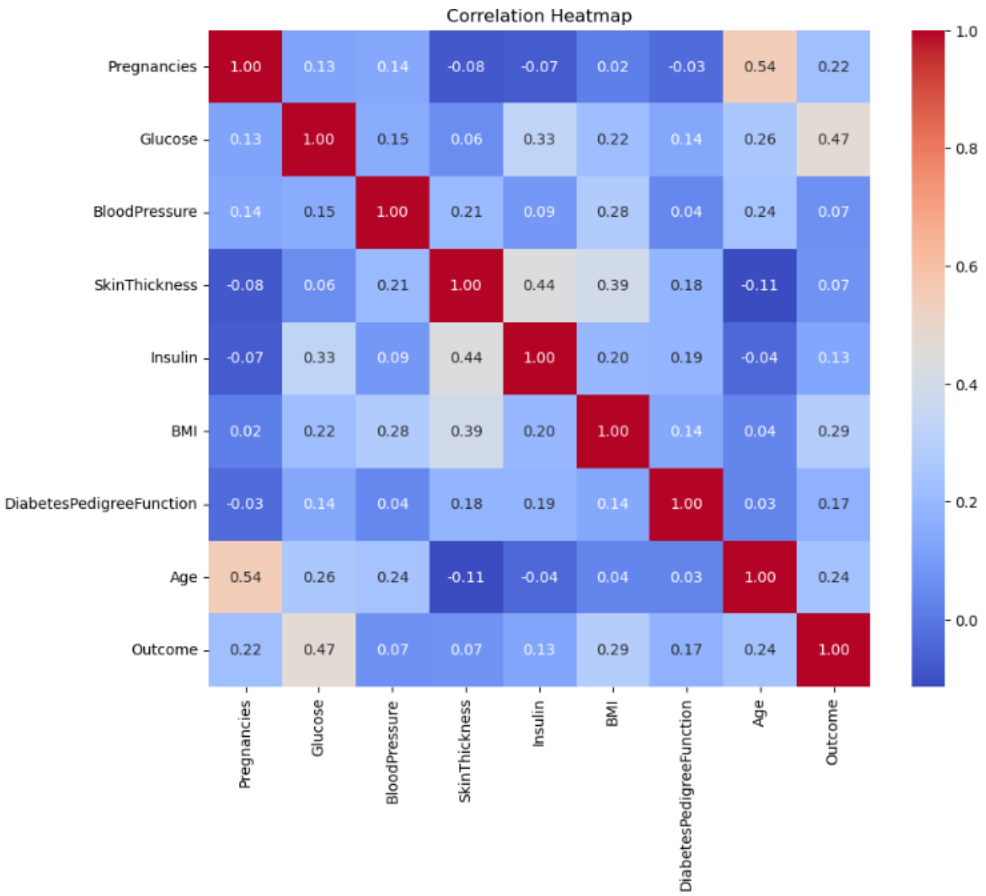

Figure 4: Correlation Heatmap of Diabetes Dataset

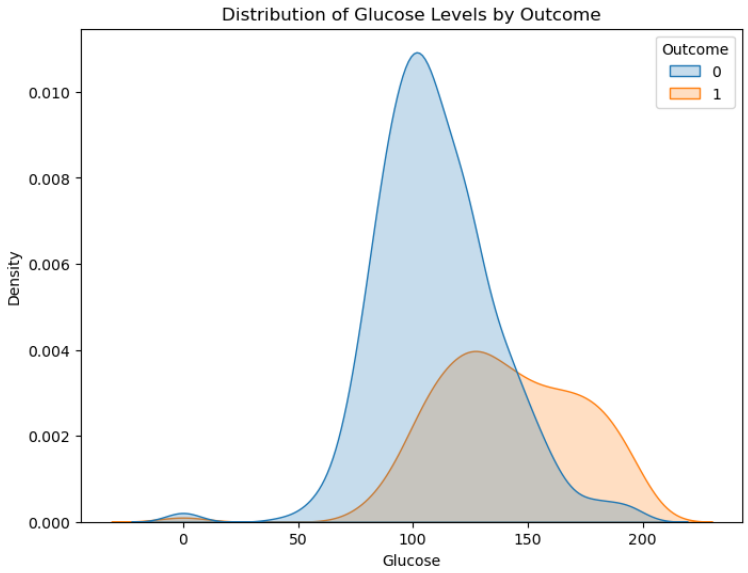

Figure 5: Distribution of Glucose Levels by Diabetes Outcome