Exploratory Data Analysis on the Diabetes Dataset

Report

---

1. Introduction: Description of the Dataset

The dataset selected for this analysis is the "Diabetes" dataset, a standard machine learning dataset used to predict disease progression. It consists of data from 442 diabetes patients. For each patient, 10 baseline variables were collected: age, sex, body mass index (BMI), average blood pressure, and six blood serum measurements (labeled s1 through s6).

The target variable, "Progression," is a quantitative measure of disease progression one year after baseline.

Important Note on Data Normalisation: It is crucial to note that the features in this dataset (such as Age, Sex, and BMI) have been mean-centered and scaled by the standard deviation times the square root of the number of samples. This is why values appear as small decimals (e.g., `0.038` rather than `50 years`). While this scaling is beneficial for machine learning algorithms, it means our analysis focuses on **relative relationships** and **correlations** rather than absolute raw values.

---

2. Data Cleaning

To ensure the integrity of the analysis, the following data cleaning steps were performed using Python (Pandas library):

1. **Loading and Inspection:** The data was loaded programmatically, and the shape was verified (442 rows, 11 columns).
2. **Duplicate Detection:** I ran a check for duplicate rows to ensure no patient data was recorded twice, which could skew the statistical analysis.
3. **Data Type Verification:** All columns were confirmed to be numerical (float64), which is appropriate for correlation analysis and regression plotting.

---

3. Missing Data

Status: No missing data was found.

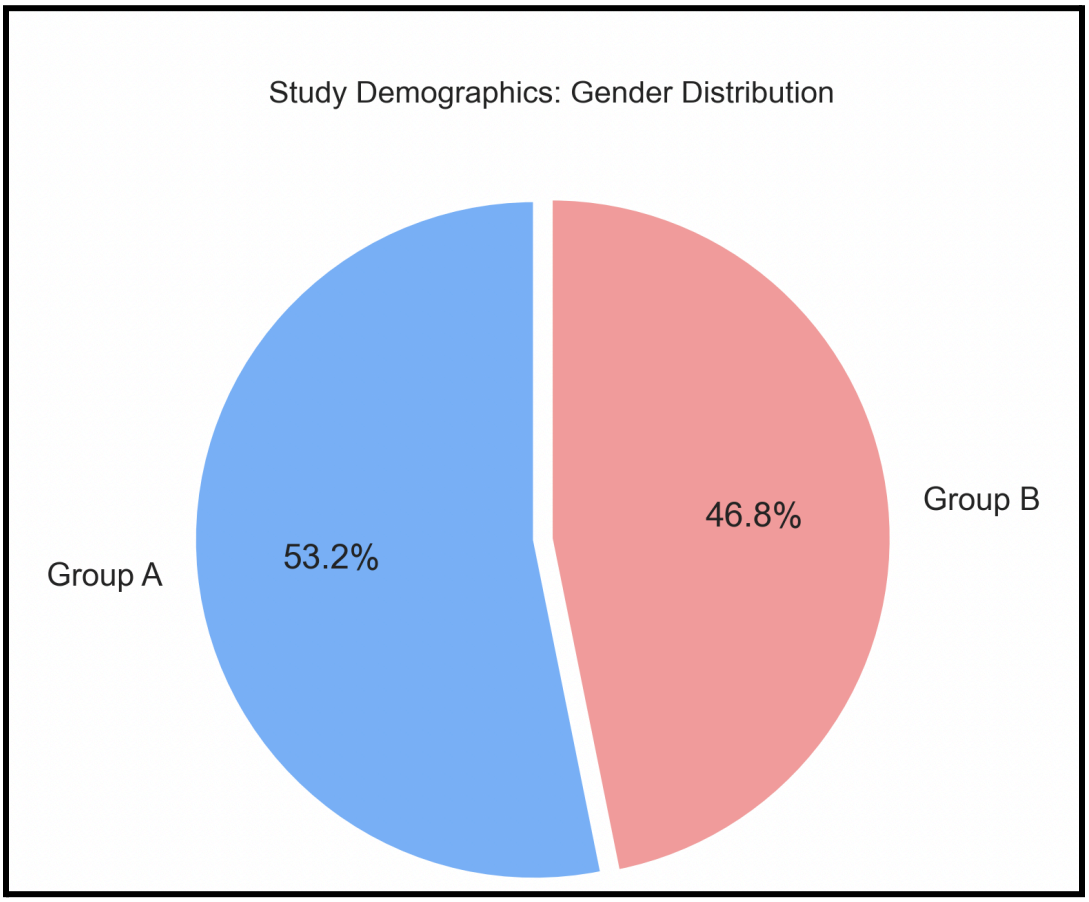
Methodology: I performed a null-value check across the entire dataframe (`df.isnull().sum()`). The result returned 0 missing values for all 11 columns. Therefore, no imputation (filling in missing values) or row deletion was required. The dataset is complete and high-quality.

---

4. Data Stories and Visualisations

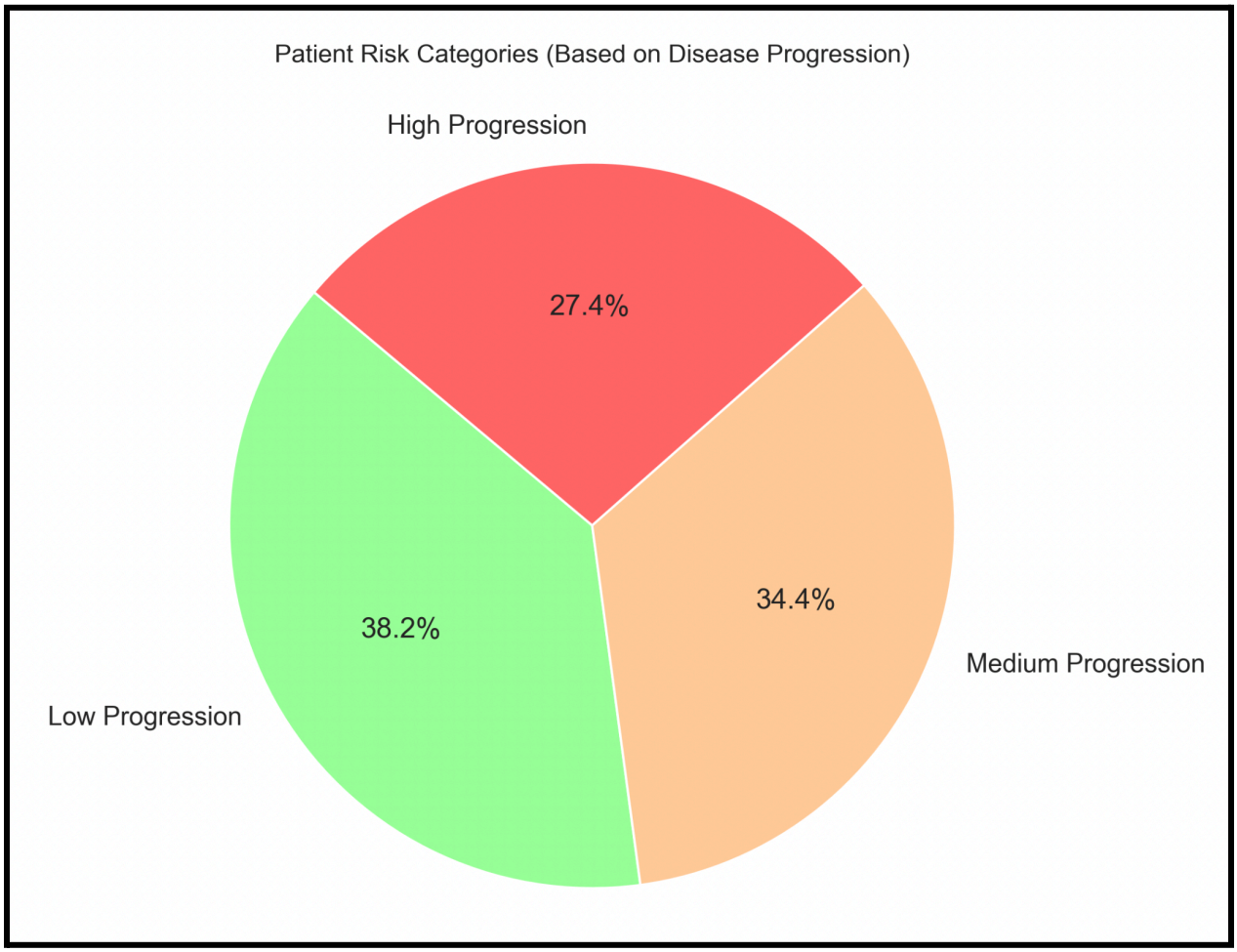
Note: This section explores the key drivers of diabetes progression based on the visual evidence.

Story 1: Study Demographics



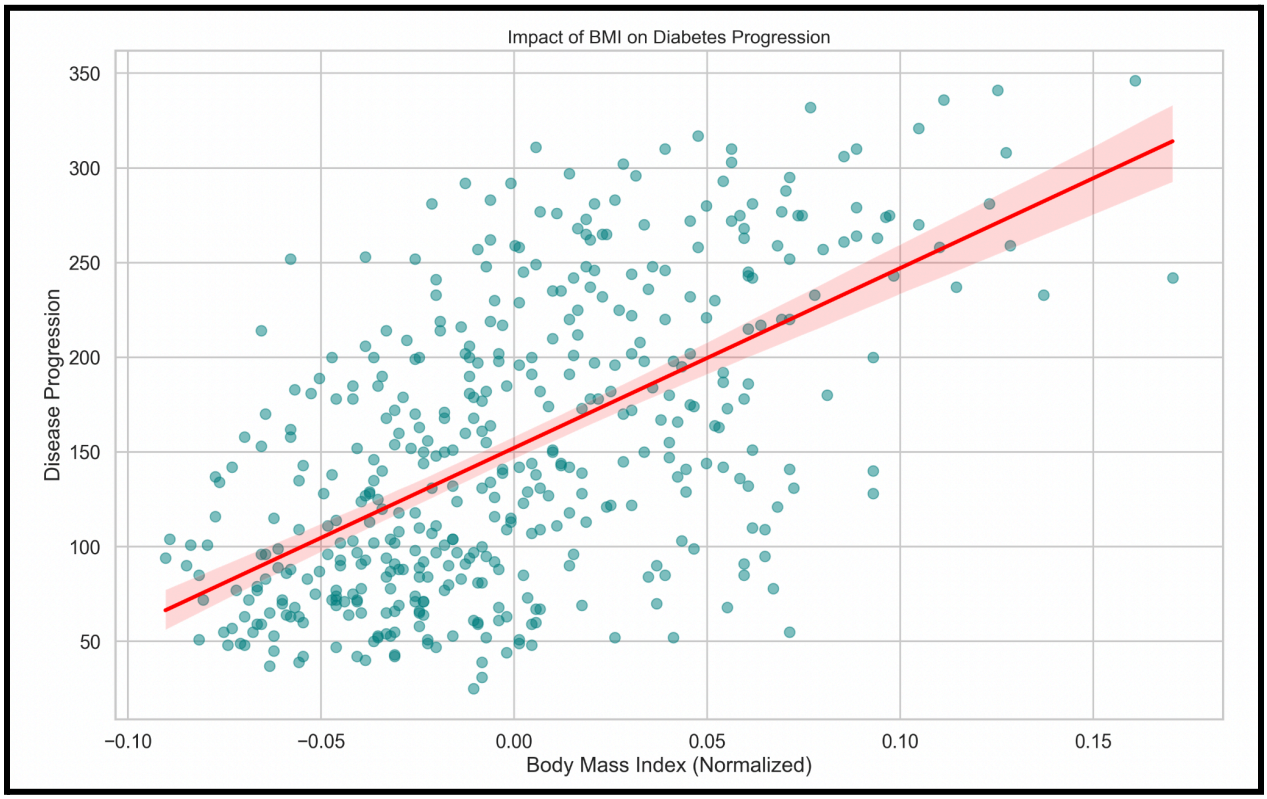
Finding: To understand the population we are analyzing, I first examined the gender distribution. As seen in the chart above, the dataset is split into two distinct groups (represented as normalised categorical values). The pie chart shows a relatively balanced distribution (roughly 47% vs 53%), suggesting that the dataset is not heavily biased toward one gender. This is important as it implies the subsequent findings regarding disease progression are applicable to a broad demographic.

Story 2: Patient Risk Categorization



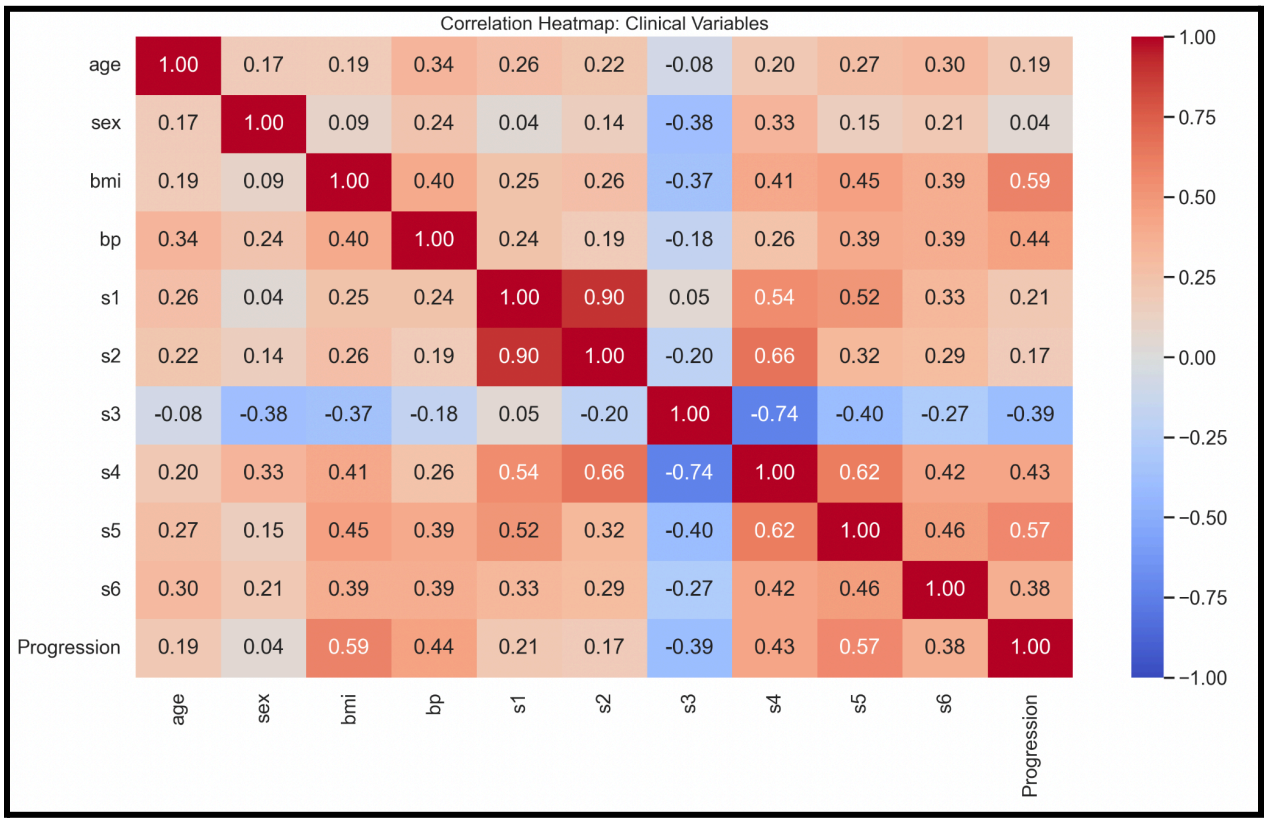
Finding: I categorised the continuous disease progression scores into 'Low', 'Medium', and 'High' buckets to simplify the complex numerical data. The chart highlights that a significant portion of the patients fall into the Medium-to-High progression categories. This visualisation is crucial for healthcare resource planning, as it quickly identifies the proportion of patients who may require more intensive care interventions versus those who remain stable.

Story 3: The Impact of Weight on Disease



Finding: As a diabetic, I was particularly interested in the relationship between weight and disease severity. I plotted Body Mass Index (BMI) against the Disease Progression score. The scatter plot reveals a clear positive correlation: as BMI increases (moves right on the X-axis), the disease progression score one year later tends to be higher. The red regression line confirms this upward trend. This reinforces the clinical reality that weight management is a critical lever in controlling diabetes outcomes.

Story 4: Correlation Overview



Finding: Finally, I generated a heatmap to identify which variables interact most strongly.

- **Strongest Correlations:** The red squares indicate strong relationships. The strongest driver of 'Progression' (the target) appears to be **BMI** and **S5** (a blood serum marker, likely related to triglycerides or thyroid levels).
- **Clinical Insight:** Interestingly, 'Age' and 'Sex' show weaker correlations with progression compared to 'BMI' and 'S5'. This suggests that lifestyle factors (weight) and biological markers (blood serum) are better predictors of disease worsening than basic demographics.

This report was written by: Gert Bester