Understanding Diabetes Risk Factors

Overview

Pima Indian Diabetes dataset is a collection of medical data from female Pima Indians that can be used to predict diabetes. It's a popular dataset used to develop machine learning models for diabetes classification

This dataset consists of **768** observations and **9** numerical features, with a binary outcome indicating whether an individual has diabetes (1: Diabetic, 0: Non-Diabetic). The analysis focused on understanding data distributions, feature correlations, and key insights regarding diabetes risk factors.

---

1. Data Quality & Cleaning

Before diving into the analysis, a few data quality issues were identified:

- Certain features, such as Glucose, Blood Pressure, Skin Thickness, and BMI, contained zero values, which are unrealistic in medical data. These were treated as missing values and either removed or imputed.
- No missing values were originally reported, but after correcting the zero values, the dataset required some refinements.
- A boxplot analysis confirmed that most extreme values were recording errors rather than true outliers.

Visualization Used: Boxplots for each feature to check for outliers and inconsistencies.

---

2. Diabetes Prevalence

- The dataset is imbalanced, with approximately 65% non-diabetic individuals (500) and 35% diabetic individuals (268).
- This imbalance suggests that special consideration is needed when building predictive models to ensure fairness and accuracy.

Visualization Used: Pie chart showing the proportion of diabetic vs. non-diabetic individuals.

---

3. Feature Comparisons with Diabetes Outcome

Each feature was compared against the diabetes outcome to understand its impact.

1. Pregnancy Trends & Diabetes

i. **Lower pregnancy counts (0-6)** are far more common in non-diabetic individuals.
ii. **Higher pregnancy counts (7+)** are more frequently observed in diabetic individuals.

This suggests that women with **more pregnancies may have a higher risk** of developing diabetes.

2. Glucose Levels & Diabetes

i. **Non-diabetic individuals:** Glucose levels primarily range between **80-120 mg/dL**.
ii. **Diabetic individuals:** More than **50% exceed 150 mg/dL**, confirming high glucose as a strong diabetes indicator.

However, there is **overlap around 130 mg/dL**, which means glucose alone isn't a definitive predictor.

3. Insulin Levels & Diabetes

i. **Diabetic individuals tend to have higher insulin levels**, but with significant variation.

The correlation between insulin and diabetes is weaker, likely due to inconsistencies in data collection.

4. Skin Thickness & Diabetes

i. **Higher skin thickness values are more common in diabetic individuals**, but there is substantial overlap with non-diabetics.

This feature alone is not a strong predictor.

5. Diabetes Pedigree Function (DPF) & Diabetes

i. Individuals with a **higher DPF score (above 0.5)** have a greater likelihood of being diabetic.

However, a few non-diabetic individuals also exhibit high DPF values.

6. Blood Pressure & BMI Analysis

i. Follows a **normal distribution**, mostly between **60-90 mmHg**.

ii. Non-diabetic individuals show a **higher concentration in the mid-range (70-80 mmHg)**.

The issue of **zero values were corrected**, leading to more reliable insights.

7. BMI (Body Mass Index):

i. BMI is **higher among diabetic individuals**, with many exceeding **30 kg/m² (obese category)**.

ii. Non-diabetic individuals have a **broader BMI range**, with a significant portion below **25 kg/m² (normal weight)**.

This reinforces the known medical fact that **higher BMI is a risk factor for diabetes**.

8. Age & Diabetes

i. **Older individuals (40+)** have a significantly higher prevalence of diabetes.

ii. **Younger individuals (<30)** are predominantly non-diabetic.

Visualization Used: Age distribution plot comparing diabetic vs. non-diabetic individuals.

4. Overall Insights & Conclusion

1. Key Risk Factors for Diabetes:

i. **High glucose levels (>150 mg/dL)**

ii. **BMI above 30 kg/m²**

iii. **Higher pregnancy counts (7+)**

iv. **Older age (correlation of 0.22)**

v. **Higher Diabetes Pedigree Function scores (>0.5)**

vi. **No single feature can predict diabetes alone**—a combination of factors is needed.

vii. **Data cleaning significantly improved the reliability** of insights, especially by correcting zero values.

Final Thought: This analysis provides a strong foundation for predictive modeling, helping to build a machine learning model for diabetes prediction. Future steps could involve feature engineering, balancing the dataset, and testing different classification models.