Diabetes Prediction using Exploratory Data Analysis

By: *Aditya Hasteer Sharma*

In this analysis, I explored various features of a dataset containing medical information to understand their relationship with diabetes. The dataset includes features such as glucose levels, insulin levels, BMI, number of pregnancies, and others. My primary objective was to predict whether an individual has diabetes based on these features. I used a combination of exploratory data analysis (EDA) and a rule-based approach to make predictions.

Approach

I focused on the relationship between glucose levels and insulin levels to predict diabetes. This decision was guided by medical knowledge indicating that high glucose levels are a strong indicator of diabetes and that insulin plays a crucial role in regulating glucose metabolism.

Steps and Workflow

1. **Setting Up the Environment**:
    - Created a conda environment to ensure a consistent and isolated environment for the analysis.
    - Installed necessary libraries such as pandas, matplotlib, and seaborn.
2. **Loading and Inspecting the Data**:
    - Loaded the dataset into a pandas DataFrame.
    - Inspected the data for missing values and calculated basic statistics to understand the data distribution.
3. **Exploratory Data Analysis (EDA)**:
    - Created scatter plots to visualize relationships between various features and the outcome variable (diabetes status).

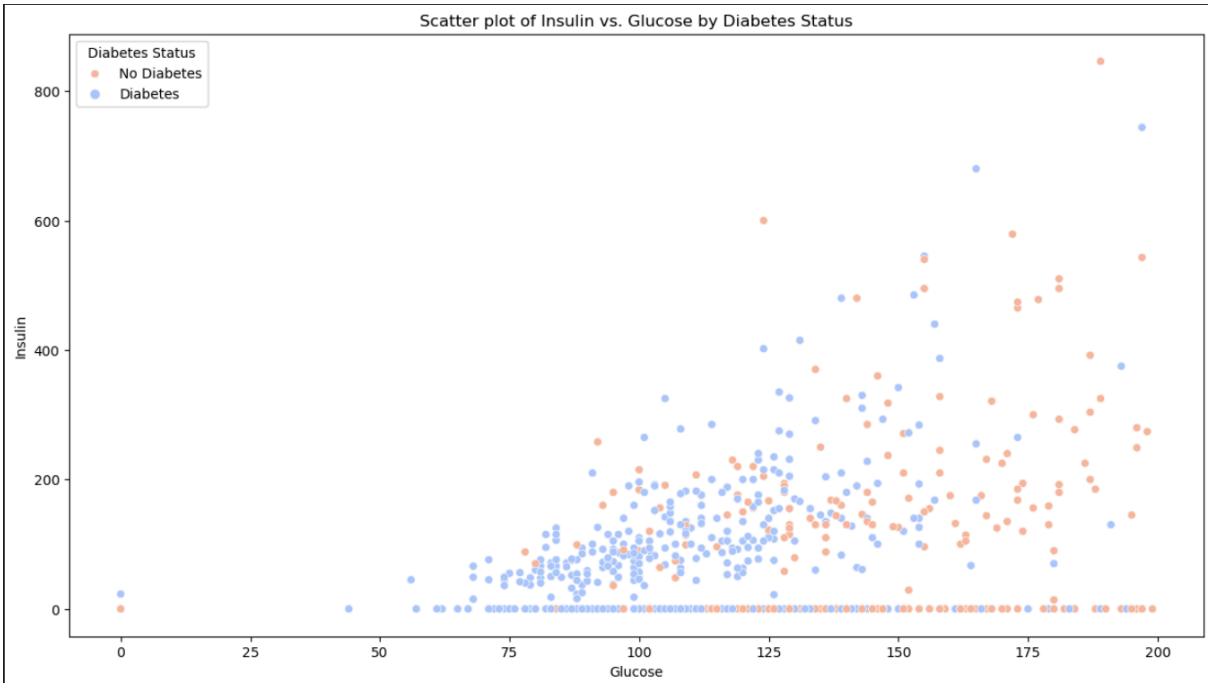
      NOTE: I took some reference from the internet to write code for specific things. Eg: I used ‘kde=True’ to form a continuous curve for better interpretation of the data.

    - Generated box plots to compare distributions of features like BMI, glucose, and insulin levels across diabetic and non-diabetic individuals.
4. **Prediction Based on Glucose and Insulin Levels**:
    - Developed a simple rule-based method for prediction: if an individual's glucose level is higher than their insulin level, they were predicted to have diabetes; otherwise, they were predicted to be non-diabetic.
    - Iterated over the dataset to apply this rule and print the results for each individual.
5. **Visualization**:
    - Created various visualizations including scatter plots, box plots, and line plots to illustrate relationships and distributions.

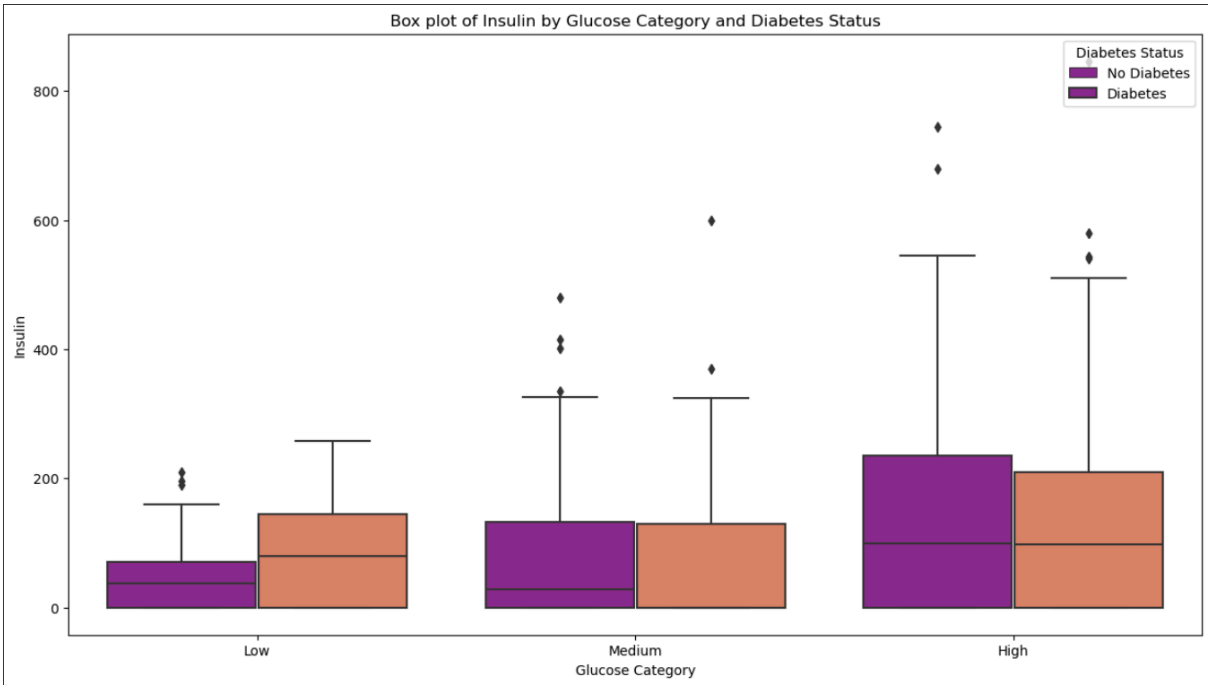
- These visualizations helped in understanding and explaining the data better.

Visualizations

- **Scatter Plot**: Showed the relationship between insulin and glucose levels, colored by diabetes status. This plot helped visualize the separation between diabetic and non-diabetic individuals based on these two features.



- **Box Plot**: Compared insulin levels across different categories of glucose levels (low, medium, high) and diabetes status. This helped in understanding the distribution andl tendency of insulin levels for different glucose categories.



Conclusion

By combining EDA and a rule-based approach, I created a straightforward and interpretable method to predict diabetes based on glucose and insulin levels. This method leveraged the strong relationship between these two features, supported by medical knowledge that I researched on the internet. The visualizations provided clear insights into the data distribution and relationships, making the analysis clear and informative.

Using a conda environment ensured that all dependencies were managed effectively, providing a reproducible and consistent setup for the analysis. This structured approach not only facilitated the analysis but also made it easier to share results, replicate the results and to describe these results to any individual.