Parth Sunil Pawar (23116073)

Exploratory Data Analysis (EDA) on PIMA Diabetes Dataset

1. Introduction

The PIMA Diabetes dataset consists of medical diagnostic measurements used to predict diabetes in patients. The dataset includes various numerical variables such as glucose levels, blood pressure, BMI, and insulin levels. The objective of this analysis is to clean the dataset, perform exploratory data analysis (EDA), and extract meaningful insights from the relationships between different variables.

2. Data Cleaning

Before conducting EDA, data cleaning was performed to ensure reliability. The following steps were taken:

- Missing Value Handling: Zero values in features such as Glucose, Blood Pressure, Skin Thickness, Insulin, and BMI were replaced with NaN since zero is not a valid measurement in these cases.

- Imputation: The missing values were replaced with the median of each respective column to retain the distribution of the dataset.

Insights:

- Some columns, especially Insulin and Skin Thickness, had multiple missing values. This indicates possible inconsistencies in data collection.

- Imputing with the median ensures that extreme values (outliers) do not distort the data.

3. Univariate Analysis

We analyzed the distribution of individual features using histograms and boxplots.

- BMI Distribution: The histogram shows that most individuals have a BMI between 20 and 40, with some outliers indicating extreme obesity.

- Age Distribution: The dataset primarily consists of adults aged between 20-80, with a higher concentration in the 30-50 range.

- Blood Pressure Boxplot: The boxplot highlights a few outliers, but most values fall within a healthy range.

Insights:

- BMI and age seem to have a reasonably normal distribution, suggesting their usability in predictive modeling.

- Blood Pressure outliers suggest possible errors in data entry or extreme medical conditions.

4. Bivariate Analysis

To analyze relationships between two variables, we plotted:

- Glucose vs. Diabetes Outcome (Boxplot): Higher glucose levels were significantly associated with diabetes.

- BMI vs. Diabetes Outcome (Histogram): Higher BMI is more common among diabetic patients.

- Glucose vs. Diabetes Outcome (Violin Plot): The violin plot showed a higher density of diabetic individuals having glucose levels above 125.

Insights:

- Glucose is a strong predictor of diabetes since most diabetic individuals have high glucose values.

- BMI also plays a role in diabetes, as overweight individuals are more likely to be diabetic.

5. Multivariate Analysis

To study relationships between multiple variables, we used:

- Correlation Heatmap: Glucose showed the highest correlation with diabetes, followed by BMI and Age.

- Pairplot: Visualized interactions between features and confirmed that diabetics tend to have higher glucose and BMI values.

Insights:

- BMI, Age, and Glucose are key indicators of diabetes.

- Other variables like Blood Pressure and Insulin show weaker correlations.

6. Age Group Analysis

We categorized individuals into different age groups and analyzed diabetes prevalence:

- Countplot of Diabetes Cases Across Age Groups: Diabetes cases are more frequent in individuals aged 40 and above.

- KDE Plot for Glucose Levels: Showed that glucose levels are generally higher in diabetic individuals across all age groups.

Insights:

- Middle-aged and older individuals (40+) are more prone to diabetes.

- Higher glucose levels are consistently associated with diabetes regardless of age.

7. Conclusion

This analysis revealed several key patterns in the PIMA Diabetes dataset:

- Glucose is the most significant factor influencing diabetes.

- Higher BMI and Age also contribute to diabetes risk.

- Diabetes cases are more prevalent in individuals aged 40+.

- Multivariate relationships confirm that glucose and BMI together provide strong predictive power.

8. Future Recommendations

- Further data collection should ensure complete and accurate insulin measurements.

- Advanced machine learning models can be built using these insights to predict diabetes with high accuracy.

- Lifestyle factors such as diet and exercise could be included to improve predictive modeling.