**Analysis of Patients with Diabetes**

**Overview**

This project provides a detailed analysis of a diabetes dataset using Python. The analysis aims to understand the relationships between various medical attributes and diabetes outcomes. Key steps include data cleaning, statistical analysis, feature engineering, and predictive modeling using logistic regression.

**Table of Contents**

1. Requirements
2. Dataset Information
3. Steps in the Analysis
   * Data Preprocessing
   * Exploratory Data Analysis (EDA)
   * Statistical Analysis
   * Feature Engineering
   * Predictive Modeling
4. Visualizations
5. Reports and Insights
6. How to Run
7. Conclusion

**Requirements**

The following Python libraries are required to run the analysis:

* pandas
* numpy
* seaborn
* matplotlib
* scipy
* sklearn

Install them using:

pip install pandas numpy seaborn matplotlib scipy scikit-learn

**Dataset Information**

The dataset is located at:

C:/Users/Admin/Downloads/health care diabetes.csv

**Attributes:**

1. **Pregnancies**: Number of pregnancies.
2. **Glucose**: Plasma glucose concentration.
3. **BloodPressure**: Diastolic blood pressure (mm Hg).
4. **SkinThickness**: Triceps skinfold thickness (mm).
5. **Insulin**: 2-Hour serum insulin (µU/ml).
6. **BMI**: Body mass index.
7. **DiabetesPedigreeFunction**: A function indicating the likelihood of diabetes based on family history.
8. **Age**: Age of the patient.
9. **Outcome**: Class variable (0 = No Diabetes, 1 = Diabetes).

**Steps in the Analysis**

**1. Data Preprocessing**

* **Null Value Handling**: Identified and handled missing values by replacing them with the median.
* **Zero Value Replacement**: Columns with invalid zero values (e.g., Glucose, BMI) were treated by replacing zeros with the median.
* **Outlier Detection and Handling**: Used box plots to detect outliers and capped extreme values to reduce skewness.

**2. Exploratory Data Analysis (EDA)**

* **Feature Distributions**: Visualized histograms and density plots for numerical attributes.
* **Correlation Analysis**: Heatmaps showed relationships between variables.

**3. Statistical Analysis**

Performed t-tests and chi-square tests to determine significant associations between features and the outcome variable.

**4. Feature Engineering**

* **BMI Category**: Categorized BMI into ‘Underweight’, ‘Normal’, ‘Overweight’, and ‘Obese’.
* **Age Group**: Grouped ages into ranges such as ‘20-30’, ‘30-40’, etc.

**5. Predictive Modeling**

Built a logistic regression model to predict diabetes outcomes:

* **Metrics**: Precision, Recall, F1-score, Accuracy.
* **Visualization**: ROC Curve and Precision-Recall Curve.

**Visualizations**

1. **Box Plots**: Displayed outliers in numerical features.
2. **Histograms and KDE**: Illustrated the distribution of key features like Glucose, BMI, and BloodPressure.
3. **Correlation Heatmap**: Highlighted relationships between variables.
4. **Pair Plots**: Showed pairwise relationships grouped by outcome.
5. **Count Plots**: Explored diabetes prevalence by age group and BMI category.

**Reports and Insights**

* **Glucose Levels**: Higher glucose values are strongly associated with positive diabetes outcomes.
* **BMI**: Obesity is a significant predictor of diabetes.
* **Age Groups**: Younger age groups (20-30) have higher non-diabetic cases, while older groups show more balanced outcomes.
* **Genetic Influence**: Higher Diabetes Pedigree Function values correlate with positive outcomes, suggesting genetic predisposition.

**How to Run**

1. Load the dataset using the provided file path.
2. Follow the steps in the code to:
   * Preprocess the data.
   * Perform EDA and statistical tests.
   * Visualize the results.
   * Train and evaluate the logistic regression model.
3. Use the interactive test section to input patient data and predict diabetes outcomes.

**Conclusion**

This project highlights the importance of data preprocessing and feature engineering in healthcare analytics. The analysis identifies key factors like glucose levels, BMI, and genetic predisposition as critical predictors of diabetes, providing valuable insights for medical decision-making.