DIABETES MILLITES

Data Preparation

- **Data Loading and Exploration**: The dataset is loaded using pandas. Initial exploratory data analysis (EDA) includes viewing the dataset with functions like head(), info(), and describe(), and checking its shape.

Data Cleaning and Preprocessing

- **Handling Missing Values**: Missing values are identified using isnull() and isnull().sum(). The missing data is handled by filling with either the median or mean, depending on the skewness of the distribution.
- **Removing Duplicates**: Duplicate entries in the dataset are removed using drop_duplicates().

Feature Engineering

- **Correlation Analysis**: A heatmap is generated using seaborn to visualize correlations between numerical features.
- **Encoding Categorical Variables**: Categorical variables are encoded using methods like LabelEncoder and get_dummies to prepare the data for modeling.
- **Feature Selection**: The Recursive Feature Elimination (RFE) method is applied with LogisticRegression to select the most relevant features for the model.

Model Building

- **Model Training**: The dataset is split into training and testing sets using train_test_split. A logistic regression model (LogisticRegression) is then trained on the training data.
- **Prediction**: The model predicts the test set outcomes and is also used for making predictions on specific input values.

Model Evaluation

- **Performance Metrics**: The model's performance is evaluated using metrics such as accuracy (accuracy_score) and the confusion matrix (confusion_matrix).

Visualization

- **Data Distribution and Relationships**: Various plots, including histograms, density plots, box plots, and pair plots, are used to visualize the distribution of individual features and relationships between features.
- **Model Performance Visualization**: A regression plot is created to visualize the relationship between actual and predicted values.

Conclusion

- The notebook effectively demonstrates the steps involved in preparing data, engineering features, building a logistic regression model, and evaluating its performance for predicting diabetes.