Diabetes Prediction Model (VAAICH NOOKK FOR VIVA PURPOSE AANE)

Overview:
This project is a Machine Learning pipeline designed to predict the likelihood of diabetes in individuals using a dataset (typically the Pima Indians Diabetes Dataset). The notebook performs data preprocessing, model training, and evaluation using Logistic Regression.

Step-by-Step Explanation:

1. Library Installation:
The first step installs essential Python libraries: pandas, numpy, matplotlib, seaborn, and scikit-learn. These are used for data analysis, visualization, and machine learning.

2. Importing Libraries:
After installation, the required libraries are imported to handle dataframes, visualize data, and train models.

3. Loading Dataset:
The dataset (CSV file) is loaded using pandas. The notebook displays the first few rows, a statistical summary, and checks for missing values. A heatmap is generated using seaborn to visualize correlations among features.

4. Data Cleaning and Preprocessing:
Certain columns like Glucose, BloodPressure, SkinThickness, Insulin, and BMI are cleaned by replacing zeros with NaN values. These missing values are then replaced (imputed) with the mean value of each column using scikit-learn’s SimpleImputer. Next, the dataset is standardized (scaled) so that each feature has a mean of 0 and standard deviation of 1, improving model performance.

5. Splitting the Dataset:
The data is split into training (80%) and testing (20%) sets using train_test_split(). This ensures that the model can be trained on one portion and tested on unseen data for fair evaluation.

6. Model Training and Evaluation:
A Logistic Regression model is trained on the training set (X_train, y_train). After training, predictions are made on the test set (X_test). The model’s performance is evaluated using accuracy score and classification report, which display precision, recall, F1-score, and overall accuracy.

7. Model Saving:
Once trained, the model is saved as a binary file 'diabetes_model.pkl' using the joblib library. This allows reusing the trained model later without retraining it.

8. Summary:
This notebook automates the complete process of preparing data, training a model, evaluating it, and saving it for deployment. It can be used as the foundation for building a diabetes prediction web or mobile application.

Usage:
1. Place your dataset CSV file in the correct path.
2. Run each cell in order from top to bottom.
3. Review the printed evaluation metrics for performance.
4. Use the saved model file to make predictions on new data.

End of Document.NANNI ONDEE