FUNDAMENTALS OF MACHINE LEARNING

MINI PROJECT

REGISTER NO : 241501103

AIM :

The aim of this project is to develop a machine learning model that can accurately predict whether a patient is likely to have diabetes based on medical attributes such as glucose level, blood pressure, BMI, age, and other factors.
Four supervised classification algorithms—**Logistic Regression, K-Nearest Neighbors (KNN), Decision Tree, and Random Forest**—are implemented, compared, and analyzed to determine the best performing model.

ALGORITHM :

1.LOGISTIC REGRESSION :

Steps :

1. Load and explore the diabetes dataset.
2. Handle missing values and perform preprocessing.
3. Split the dataset into training and testing sets.
4. Normalize/scale numerical features (important for LR).
5. Train the Logistic Regression model on the training data.
6. Use the model to predict diabetes on the test data.
7. Evaluate performance using accuracy, precision, recall, F1-score, and confusion matrix.

2.K-NEAREST-NEIGHBOUR:

Steps :

1. Load and preprocess the dataset.
2. Normalize all features (mandatory for KNN).
3. Choose the value of **k** (commonly odd values: 3, 5, 7…).
4. For each test sample:
   - Calculate distance (Euclidean) from all training samples.
   - Identify the **k nearest neighbors**.
   - Assign the class that appears most frequently.
5. Evaluate the model using test data metrics.

3.RANDOM FOREST CLASSIFIER :

Steps :

1. Load and preprocess the dataset.
2. Split into training and testing data.
3. Initialize Random Forest with a chosen number of trees (e.g., 100).
4. Train the model on the training dataset.
5. Combine predictions from all trees using majority voting.
6. Predict the outcome for test data.
7. Evaluate the model using accuracy and other metrics.

4. DECISION TREE CLASSIFIER :

Steps :

1. Import and preprocess the diabetes dataset.
2. Split the data into train and test sets.
3. Initialize the Decision Tree model (criterion = Gini/Entropy).
4. Train the model on the training dataset.
5. Predict the class labels for the test dataset.
6. Visualize the tree (optional).
7. Evaluate model performance.

RESULT :

The diabetes prediction model was developed using four machine learning algorithms: **Logistic Regression, K-Nearest Neighbors (KNN), Random Forest, and Decision Tree**. After training and testing each model on the dataset, the following accuracies were obtained:

Algorithm	**Accuracy**
Logistic Regression	**0.85**
K-Nearest Neighbors	**0.84**
Random Forest	**0.85**
Decision Tree	**0.85**