K-Nearest Neighbors (KNN)

**Aim**

To develop a predictive model using the K-Nearest Neighbors (KNN) algorithm for assessing diabetes risk based on the Pima Indians Diabetes Dataset. This model aims to provide accurate predictions to assist healthcare professionals in early diabetes detection and intervention.

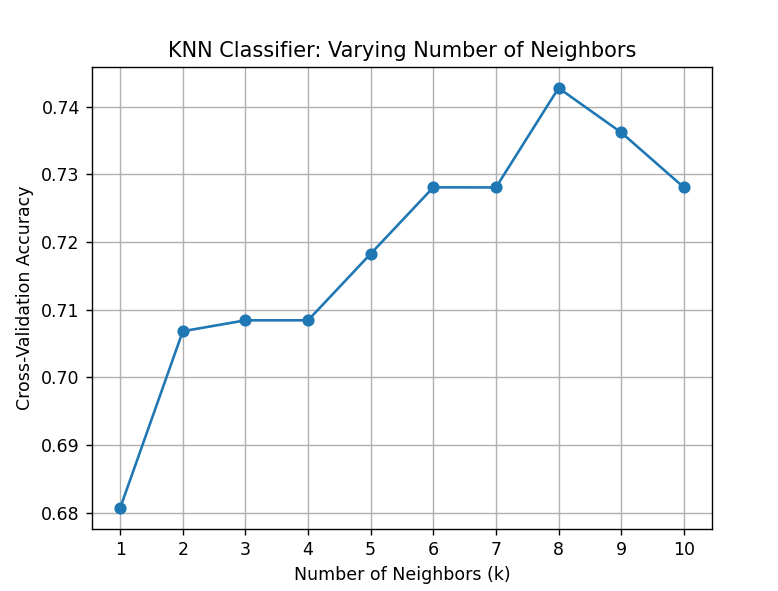
**Problem Statement**

Diabetes is a significant health concern that affects millions of individuals worldwide. Early detection is crucial for managing the disease and reducing the risk of complications. This project aims to create a KNN classification model that can accurately predict whether an individual has diabetes based on various health metrics, enabling timely intervention and better health outcomes.

**Procedure (Detailed)**

* **Data Collection**:
  + The Pima Indians Diabetes Dataset is sourced from a public repository, containing health metrics for female patients.
* **Data Preprocessing**:
  + **Imputation**: Missing values in the dataset are filled using the mean strategy to ensure a complete dataset for analysis.
* **Feature Selection**:
  + The features selected for the model include:
    - Pregnancies
    - Glucose
    - BloodPressure
    - SkinThickness
    - Insulin
    - BMI
    - DiabetesPedigreeFunction
    - Age
  + The target variable is the presence of diabetes (0 = No Diabetes, 1 = Diabetes).
* **Data Splitting**:
  + The dataset is split into training (80%) and testing (20%) sets to evaluate the model's performance.
* **Model Training**:
  + A loop is created to evaluate the KNN classifier with varying values of kkk (from 1 to 10) using cross-validation to determine the best value based on accuracy.
* **Model Evaluation**:
  + The best kkk value is selected based on the highest accuracy, and the model is retrained using this optimal kkk.
  + The model's performance is assessed on the test dataset, and accuracy is calculated.
* **User Input for Prediction**:
  + A function is implemented to collect user input for prediction, which allows the model to make predictions based on user-provided health metrics.
* **Visualization**:
  + A scatter plot visualizes the original data points and the new data point, showcasing how the model classifies the user input in relation to existing data.

OUTPUT

  
The best k value is 8 with an accuracy of 0.7428

Accuracy of KNN with k=8: 0.7468

**Inference**

From a business perspective, implementing a KNN-based predictive model for diabetes risk assessment can greatly enhance the healthcare sector's ability to provide preventive care. By identifying individuals at high risk for diabetes through timely predictions, healthcare providers can initiate early interventions, reducing the incidence of diabetes-related complications and associated costs. This proactive approach not only improves patient outcomes but also optimizes resource allocation within healthcare systems. Additionally, engaging patients with personalized assessments fosters awareness and motivates lifestyle changes, further aiding in diabetes prevention. Overall, this model serves as a strategic asset in promoting public health initiatives and improving the efficiency of diabetes management programs.