KNN classification:

K	**Test Accuracy**
1	96.667 %
3	96.667%
5	96.667%
10	93.333%
15	96.667%

Best Test Accuracy: **__96.667%__** (K= 1, 3, 5, 15)

KNN regression:

K	**Test Mean Squared Error**
1	5984.876404494382
3	4280.6379525593
5	3396.3928089887636
10	3428.6803370786515
15	3583.8870911360796

Best Test Mean Squared Error: __3396.3928089887636__ (K= 5)

Introduction to the Problem:

Following the implementation of KNN classification on the Iris dataset, the focus shifted to regression analysis using the diabetes dataset. Unlike the discrete classifications required for the Iris dataset, this task involved predicting continuous values to measure the progression of diabetes. This presented a different challenge and required adjustments to the methodology.

Overview:

The Iris dataset contained 150 data points representing three types of flowers, while the diabetes dataset comprised 442 data points with 10 features such as age, sex, BMI, and blood pressure. The latter dataset was more complex and necessitated a careful approach to ensure accurate predictions.

Step-by-Step Implementation:

Data Preprocessing:

- Loaded the data using the pandas library.
- Randomized the data to eliminate ordering bias.
- Split the dataset into 80% training and 20% testing subsets.

Distance Calculation:

- Used the Euclidean Distance formula, similar to the classification task, to measure the similarity between patient records.

Prediction Function:

- Computed the average of the disease progression values of the K nearest neighbors.
- Utilized numpy for efficient mean calculations to ensure accuracy.

Error Calculation:

Calculated the Mean Squared Error (MSE) to evaluate prediction accuracy, involving the following steps

- Determined the difference between predicted and actual values.
- Squared the differences to avoid negative deviations.
- Averaged the squared differences across all test samples.

Results Analysis:

The impact of varying K-values on prediction accuracy was as follows:

- K=1: High error (MSE = 5984.87).
- K=3: Improved results (MSE = 4280.63).
- K=5: Further improvement (MSE = 3396.39).
- K=10: Optimal performance (MSE = 3428.68).
- K=15: Slight decline in accuracy (MSE = 3583.88).

These results indicate that increasing the number of neighbors generally improves prediction accuracy up to a certain point. K=5 provided the best results, balancing noise reduction and prediction reliability. While K=10 performed nearly as well, higher values, such as K=15, showed diminishing returns.

Implementation_Challenges:

- Adapting the code to predict continuous values instead of classifications.
- Debugging the calculate_mse function to ensure correct error computation.
- Verifying the accuracy of the MSE values to validate the model's performance.

Final Observations:

The regression implementation demonstrated the ability to predict diabetes progression with reasonable accuracy, particularly when using 10 nearest neighbors. This balance allowed the model to mitigate random noise while maintaining prediction precision. Overall, the approach proved effective and underscored the importance of parameter tuning in machine learning tasks.