Iris Dataset Task

The Iris dataset is a classic dataset used in machine learning and statistics. It contains measurements for 150 Iris flowers from three different species: Setosa, Versicolour, and Virginica. The four measurements (features) included are sepal length, sepal width, petal length, and petal width. The dataset is often used for entry-level machine learning tasks because it is relatively small, doesn't have any missing values, and the features can effectively predict the flower species (class). It's great for practicing data exploration, visualization, and simple machine learning techniques.

In this task, we will first explore the data (Part 1) and then apply a simple machine learning model on it (Part 2).

For more context, visit:

- Wikipedia
- Gist

You may use the following script to help you get started

```
# import necessary libraries
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.datasets import load_iris
from sklearn.model_selection import train_test_split
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import accuracy_score
from sklearn.metrics import plot_confusion_matrix
```

Part 1 - Data Exploration and Visualisation

1. How many rows and columns does the dataset have? (i.e., what is the dimensionality of the dataset)
2. List all the column names present in the dataset.
3. Identify the data type of each column in the dataset.
4. Visualize the pairwise relationships between features, distinguishing the data points by target species.

5. Display a histogram for each feature in the dataset to see the frequency distribution of each feature.

6. Display box plots for each feature to identify potential outliers.

7. Perform a groupby operation on the target variable to find out the mean of the other variables for each class of flowers.

8. Display the correlation between all columns in a heatmap.

Reflective Questions:

1. While observing the pairplot and heatmap, what specific patterns or relationships did you notice in the dataset? How do these visualizations help you better understand the structure and relationships in the data?

2. Considering the visualizations, which feature(s) do you think are more indicative of the species of the flower? Can you identify any potential problems or limitations with using these features to predict the flower species?

Part 2 - Simple Machine Learning

1. Separate the dataset into features (X) and target variable (y). What is the shape of X and y?

2. Partition the data into training (80%) and test (20%) sets. How many samples are present in each?

3. Initialize a k-nearest neighbors classifier with k=3, train it on the training data, and then predict the species for the test data. How many predictions did it make?

4. Determine the accuracy of the classifier on the test data.

5. Try different 'k' values for the KNN classifier. Provide a visual representation of how well the classifier performed on the test data using a confusion matrix.

Reflective Questions:

1. The model appears to have a high accuracy. Can you think of reasons why the model performed so well on this particular task?

2. K-nearest neighbors (KNN) is a type of instance-based learning algorithm. Can you think of any scenarios or types of data where this algorithm might not perform as well?