Assignment 1 : Classification/Regression

1. Problem Statement: Iris Flowers Classification

The objective of this task is to classify the Iris flowers into three distinct species: Iris-setosa, Iris-versicolor, and Iris-virginica. This classification will be based on the morphological measurements of the flowers, specifically the lengths and widths of their petals and sepals.

Steps to follow:
a. Data Acquisition: The Iris dataset, which is publicly available, can be downloaded from the UCI Machine Learning Repository. The dataset consists of 150 samples with 4 features each.
b. Data Preparation: Prepare the dataset by loading it into a suitable data structure, such as a Pandas DataFrame.
c. Feature Selection: Use the provided features (sepal length, sepal width, petal length, petal width) to build a classification model.
d. Model Training: Develop a classification model to categorize the Iris flowers into the three species.
e. Model Evaluation: Evaluate the model's performance using appropriate metrics. Present the results using a confusion matrix.
f. Visualization: Plot the classification results to visually assess the performance of the model.

The aim is to achieve an accurate classification of the Iris flower species and to demonstrate the effectiveness of the model through evaluation metrics and visualizations.

Download the Iris dataset from the UCI Machine Learning Repository: Iris Dataset.

2. Problem Statement: Overfitting and Underfitting Analysis using Regression/Classification Models

In this task, you are required to perform a detailed analysis of overfitting and underfitting issues using various regression/classification models on multiple datasets. The steps to be followed are:

a. Dataset Selection and Preparation:
- Download the following datasets:
  - Mobile Price Classification Dataset
  - Housing Price Dataset
  - Melbourne Housing Snapshot Dataset
- Analyze the features of each dataset and choose the relevant attributes for prediction or classification tasks.

b. Data Preprocessing:

- Handle missing values in the datasets using appropriate imputation techniques.
- Normalize the datasets if necessary to ensure the features are on a similar scale.

c. Model Development:

- Split each dataset into training and testing sets. You may use techniques like stratified k-fold cross-validation to ensure a balanced split.
- Develop regression models (e.g., linear regression, multiple regression) or classifier models to predict the target variable.
- For each model, estimate the parameters and generate predictions on both training and testing sets.

d. Overfitting/Underfitting Analysis:

- Plot the training and validation loss curves to visualize and identify overfitting or underfitting scenarios.
- Evaluate the models using metrics such as Mean Squared Error (MSE) and R^2 score on both training and testing sets.
- Compare the performance of different models and discuss the observations related to overfitting and underfitting.

e. Reporting:

- Summarize the results and insights from your analysis.
- Highlight any patterns or trends observed during the study.
- Provide recommendations for improving model performance and addressing overfitting or underfitting issues.

By conducting this comprehensive analysis, you will gain a deeper understanding of how different regression models perform on various datasets and the common pitfalls associated with overfitting and underfitting in machine learning.