ASSIGNMENT 3

Possible Points: 100 **Date: March 17th, 2025**

Assignments should be done on an individual basis.
Download the dataset **insurance.csv** provided to perform the following tasks:

Q.1 Perform regression using SVM regression model on insurance dataset. Predict the insurance charges from the columns age, sex, bmi, children, smoker, and region. Follow the steps below: **(50 Points)**

- Check for null values in the dataset. If present, remove those values. Convert the categorical values into numerical values using label encoder. Normalize the numerical features using standard scaler. Split the data into 70-30 training testing ratio. **(10 Points)**
- Visualize the data using histogram for age, bmi and charges. Plot a scatter plot to visualize the relation between charges and bmi. **(10 points)**
- Train the model using SVM regression. Train two models one using kernel linear and one using poly. Use bootstrapped training samples. **(15 Points)**
- Evaluate the models using MSE, R squared score and bootstrap confidence interval. Which model performs better? **(15 Points)**

Q.2 Perform classification using SVM classifier model on insurance dataset. Predict the sex from the columns age, children, bmi, smoker, region, and charges. Follow the steps below: **(50 Points)**

- Check for null values in the dataset. If present, remove those values. Convert the categorical values into numerical values using label encoder. Normalize the numerical features using standard scaler. Split the data into 80-20 training testing ratio. **(10 Points)**
- Visualize the data using bar plot to analyze smoker vs sex. Plot a correlation heatmap to understand feature relationships. **(10 Points)**
- Train the model using SVM classifier. Use rbf kernel. Train the model using K-fold cross validation and LOOCV. **(20 Points)**
- Evaluate the model using accuracy score, precision, f1 score, recall. Compare performance between K-fold cv and LOOCV. **(10 Points)**