Machine Learning Assignment (20 marks)

Instructions:-Please submit the Colab Notebook as the solution to the assignment. Include any necessary assumptions in a text within the Colab notebook.

Question1: A biologist hypothesizes that there is a direct correlation between the quantity of fertilizer given to tomato plants and the resulting tomato yield.

Plant	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*
x	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5
y	3.9	4.4	5.8	6.6	7.0	7.1	7.3	7.7

For an experiment, eight tomato plants of the same variety were randomly chosen and treated weekly with a solution containing x grams of fertilizer dissolved in a fixed amount of water. The yield of tomatoes in kilograms, denoted as y, was then measured and recorded.

(a) Plot a scatter diagram of yield, y, against the amount of fertilizer, x.

(b) Calculate the equation of the least squares regression line of y on x.

(c) Estimate the yield of a plant treated, weekly, with 3.2 grams of fertilizer.

(d) Indicate why it may not be appropriate to use your equation to predict the yield of a plant treated, weekly, with 20 grams of fertilizer.

Question 2: Check the "Titanic Dataset": https://www.kaggle.com/competitions/titanic Build predictive models to predict the survival of a passenger based of the given features.

i) Logistic Regression

ii) K-Nearest Neighbour classifier

Compare the performances of the models with respect to the accuracy, recall, precision, and F-score metrics.

Question 3: Use the Housing dataset (https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques) and build a predictive model to predict the selling price of a house.

Use the following algorithms:

i) Linear Regression

ii) K-Nearest Neighbour Regressor

Perform suitable pre-processing of the training data to find out - feature importance, presence of multi-collinearity, linearity conditions. Check for the outliers in the features and if present, remove it. Apply feature scaling (verify if it affects the performance of the model). Train the models and compare the performances of the models with r-squared metric.

Question 4:Classification Problem Statement (SVM+PCA).

Objective:

The objective is to simulate a high-dimensional dataset with two classes and build a classifier using Support Vector Machines (SVM) after reducing the data dimensionality using Principal Component Analysis (PCA). Students will evaluate the classifier's performance and explore how PCA impacts accuracy.

Dataset:-

Students can use make_classification function below from sklearn.datasets to generate the data:

from sklearn.datasets import make_classification

X, y = make_classification(

```
n_samples=1000,

n_features=50,

n_informative=10,

n_redundant=10,

n_classes=2,

class_sep=1.5,

random_state=42
```

)

Tasks:

Data Generation and Visualization

Generate data using make_classification

Perform exploratory data analysis (EDA)

Visualize data using 2D PCA for initial understanding

Preprocessing:

Standardize features using StandardScaler

Dimensionality Reduction with PCA:

Apply PCA

Plot cumulative explained variance

Select number of components to preserve ~95% variance

SVM Classification:

Train-test split (e.g., 80/20)

Train an SVM classifier using both:

Linear Kernel

RBF Kernel

Use GridSearchCV to tune hyperparameters like C and gamma

Performance Evaluation:

Accuracy

Confusion Matrix

Precision, Recall, F1-score

ROC Curve and AUC score

Comparative Study:

Compare SVM performance with and without PCA

Discuss the impact of PCA on training time and accuracy

```
Start coding or generate with AI.
```