Assignment 1

Total Marks: 50

Release Date: 26 Jan 2025

Date of Submission: 2 March 2025

General Instructions (Non-conformance will result in penalties):

- A report must be written for the assignment. Please make sure that the report is neat, properly formatted, and contains all relevant information to the tasks in the assignment.
- Each question within the report shall not exceed 1 page. (Q1 in one page, Q2a and Q2b in one page and so on).
- Grading of the assignment shall be done based on both, the codes and the report. The report shall reflect your understanding regarding the concepts, methods that you have used and the analysis of the results. Please keep the content of your reports crisp and precise.
- Submission should contain the following three files - .ipynb file, .py file and report (.pdf). Please name each of these as roll_number.ipynb, roll_number.py and roll_number.pdf.
- Within the code notebook, please make proper sections for each question as discussed in the demonstration session.
- Any deviation from the above-mentioned submission format or deadline breach will result in penalty in marks – a zero will be awarded.
- Even though intellectual discussion is encouraged, any plagiarism within the report is completely unacceptable and will result in nullifying of marks and a possible action for academic misconduct (according to the relevant institute regulations).
- The penalty policy, as shared earlier, shall apply for late submissions.
- The assignment must be done in a group of **two.** Submit only 1 report per group.

Q1: Data Preprocessing – 10 Marks

You have been provided with a CSV file "Cars93.csv." The given dataset is related to cars and contains 26 columns. In the given dataset, “Price” is the target variable (i.e., the output). The marks distribution according to the tasks are as follows:

1. Assign a type to each of the following features (a) Model, (b) Type, (c) Max. Price and (d) Airbags from the following: ordinal/nominal/ratio/interval scale.
2. Write a function to handle the missing values in the dataset (e.g., any NA, NaN values).
3. Write a function to reduce noise (any error in the feature) in individual attributes.
4. Write a function to encode all the categorical features in the dataset according to the type of variable jointly.
5. Write a function to normalize / scale the features either individually or jointly.
6. Write a function to create a random split of the data into train, validation and test sets in the ratio of [70:20:10].

Q2a: Linear Regression Task. - 6 Marks

Use the “linear_regression_dataset.csv”

Implement the linear regression model to predict the dependency between two variables.

1. Implement linear regression using the inbuilt function “LinearRegression” model in sklearn.
2. Print the coefficient obtained from linear regression and plot a straight line on the scatter plot.
3. Now, implement linear regression without the use of any inbuilt function.
4. Compare the results of 1 and 3 graphically.

Q2b: Logistic Regression Task. - 4 Marks

Use the “logistic_regression_dataset.csv”

1. Split the dataset into training set and test set in the ratio of 70:30 or 80:20
2. Train the logistic regression classifier (using inbuilt function: LogisticRegression from sklearn).
3. Print the confusion matrix and accuracy.

Q3: SVM - 15 marks

Use the dataset “Bank_Personal_Loan_Modelling.csv”

1. Store the dataset in your google drive and in Colab file load the dataset from your drive.
2. Check the shape and head of the dataset.
3. Age, Experience, Income, CCAvg, Mortgage, Securities are the features and Credit Card is your Target Variable.
    i. Take any 3 features from the six features given above
    ii. Store features and targets into a separate variable
    iii. Look for missing values in the data, if any, and address them accordingly.
    iv. Plot a 3D scatter plot using Matplotlib.
4. Split the dataset into 80:20. (3 features and 1 target variable).
5. Train the model using scikit learn SVM API (LinearSVC) by setting the regularization parameter C as C = {0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000}.
    i. For each value of C Print the score on test data
    ii. Make the prediction on test data
    iii. Print confusion matrix and classification report
6. Use gridSearchCV a cross-validation technique to find the best regularization parameters (i.e.: the best value of C).

In the report provide your findings for the output generated for all the kernels used and also describe the changes that happened after changing the regularization hyperparameter.

Q4: Decision Tree and Random Forest – 15 Marks

Load the IRIS dataset. The dataset consists of 150 samples of iris flowers, each belonging to one of three species (setosa, versicolor, or virginica). Each sample includes four features: sepal length, sepal width, petal length, and petal width.

1. Visualize the distribution of each feature and the class distribution.
2. Encode the categorical target variable (species) into numerical values.
3. Split the dataset into training and testing sets (use an appropriate ratio).
4. Decision Tree Model
    i. Build a decision tree classifier using the training set.
    ii. Visualize the resulting decision tree.
    iii. Make predictions on the testing set and evaluate the model's performance using appropriate metrics (e.g., accuracy, confusion matrix).
5. Random Forest Model
    i. Build a random forest classifier using the training set.
    ii. Tune the hyperparameters (e.g., number of trees, maximum depth) if necessary.
    iii. Make predictions on the testing set and evaluate the model's performance using appropriate metrics and compare it with the decision tree model.