ENT 853: Machine Learning

Question 1: Data Loading, Cleaning, Analysis, and Visualization (20 Marks) [Compulsory]

Dataset: Titanic Survival Dataset

A. Theory Section (4 Marks)

1. **(1 mark)** Explain why handling missing data is crucial in machine learning. Provide two common techniques for dealing with missing values.

2. **(1 mark)** What is feature scaling, and why is it important in machine learning?

3. **(1 mark)** Differentiate between classification and regression in the context of supervised learning.

4. **(1 mark)** Why is data visualization important in exploratory data analysis (EDA)?

B. Practical Section (16 Marks)

1. **Data Loading (3 marks)**
   - Load the Titanic dataset using Pandas.
   - Display the first five rows and summarize the dataset.
2. **Data Cleaning (6 marks)**
   - Identify and handle missing values appropriately. (3 marks)
   - Convert categorical variables (e.g., Sex, Embarked) into numerical representations. (3 marks)
3. **Data Analysis (4 marks)**
   - Provide summary statistics of the dataset. (2 marks)
   - Analyze and display the survival rate across different passenger classes. (2 marks)
4. **Data Visualization (3 marks)**
   - Create a histogram to illustrate the distribution of passengers' ages. (1.5 marks)
   - Generate a bar plot showing the survival rate across different passenger classes. (1.5 marks)

Question 2: Regression Analysis (20 Marks)

Dataset: Boston Housing Dataset

1. **Data Preparation (6 marks)**
   - Load the dataset and prepare the features and target variable. (3 marks)

- Perform basic exploratory data analysis (summary statistics, checking for missing values). (3 marks)

2. **Model Implementation (7 marks)**
   - Implement a **Linear Regression model** using `scikit-learn`. (4 marks)
   - Train the model and display the regression coefficients. (3 marks)
3. **Model Evaluation (7 marks)**
   - Evaluate the model using **Mean Squared Error (MSE)** and **R^2 score**. (4 marks)
   - Plot the predicted vs. actual values to assess model performance. (3 marks)

Question 3: Decision Tree Classification (20 Marks)

Dataset: Mushroom Classification Dataset

1. **Data Preparation (6 marks)**
   - Load the dataset and encode categorical variables appropriately. (3 marks)
   - Split the dataset into training and testing sets. (3 marks)
2. **Model Implementation (7 marks)**
   - Implement a **Decision Tree classifier** to predict whether a mushroom is edible or poisonous. (4 marks)
   - Train the model and display the decision tree structure. (3 marks)
3. **Model Evaluation (7 marks)**
   - Evaluate the model's accuracy. (4 marks)
   - Visualize the feature importance of different mushroom characteristics. (3 marks)

Question 4: Random Forest Classification (20 Marks)

Dataset: Wine Quality Dataset

1. **Data Preparation (6 marks)**
   - Load the dataset and split it into training and testing sets. (3 marks)
   - Perform data normalization and handle class imbalance if needed. (3 marks)
2. **Model Implementation (7 marks)**
   - Implement a **Random Forest classifier** to predict wine quality. (4 marks)
   - Tune hyperparameters (e.g., number of trees, max depth) to improve performance. (3 marks)
3. **Model Evaluation (7 marks)**
   - Assess the model's performance using **accuracy, precision, recall, and F1-score**. (4 marks)
   - Plot a confusion matrix to visualize classification performance. (3 marks)

Question 5: Support Vector Machine (SVM) Classification (20 Marks)

Dataset: Iris Flower Dataset

1. **Data Preparation (6 marks)**
   - Load the dataset and preprocess it for **SVM classification**. (3 marks)
   - Perform feature scaling to improve the performance of SVM. (3 marks)
2. **Model Implementation (7 marks)**
   - Implement an **SVM classifier** to classify the iris species. (4 marks)
   - Train the model using different kernel types (linear, RBF) and compare results. (3 marks)
3. **Model Evaluation (7 marks)**
   - Evaluate the classifier's performance using accuracy and confusion matrix. (4 marks)
   - Visualize the decision boundaries of the SVM model. (3 marks)

Question 6: K-Nearest Neighbors (KNN) Classification (20 Marks)

Dataset: Breast Cancer Wisconsin Dataset

1. **Data Preparation (6 marks)**
   - Load the dataset and **normalize the feature variables**. (3 marks)
   - Split the dataset into training and testing sets. (3 marks)
2. **Model Implementation (7 marks)**
   - Implement a **K-Nearest Neighbors classifier** to predict the diagnosis (malignant or benign). (4 marks)
   - Experiment with different values of K and evaluate their impact on accuracy. (3 marks)
3. **Model Evaluation (7 marks)**
   - Assess the model's accuracy using cross-validation. (4 marks)
   - Plot the error rate as a function of K to determine the optimal K value. (3 marks)

Appendix: Data Sources (Requires internet connection and Kagle Account)

1. Titanic Survival Dataset at https://www.kaggle.com/c/titanic/data
2. Boston Housing Dataset at https://www.kaggle.com/datasets/altavish/boston-housing-dataset
3. Mushroom Classification Dataset at https://www.kaggle.com/datasets/uciml/mushroom-classification
4. Wine Quality Dataset at https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009
5. Iris Flower Dataset at https://archive.ics.uci.edu/ml/datasets/iris
6. Breast Cancer Wisconsin Dataset at https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data