

Mid Term Exam

Machine Learning

Task 1: Medical Cost Prediction using Multiple Regression

You are provided with the **Medical Cost Personal Dataset**, which contains demographic and health-related information of individuals along with their **medical insurance charges**.

Your task is to **build a Multiple Linear Regression model** to predict medical charges and evaluate its performance.

Step 1: Dataset Loading and Inspection

1. Load the dataset (`insurance.csv`) into a Pandas DataFrame.
2. Display the first five rows of the dataset.
3. Use `.info()` and `.describe()` to summarize the dataset.
4. Identify:
 - o Target variable
 - o Feature variables

Briefly comment on the type of problem and the nature of the dataset.

Step 2: Data Preprocessing

2.1 Handling Missing Values

- Check for missing or null values.
- Handle missing values appropriately (if any).
- Explain your approach.

2.2 Feature Encoding

- Encode categorical variables such as:
 - o sex
 - o smoker
 - o region
- Use **One-Hot Encoding** or **Label Encoding** where appropriate.
- Justify your encoding choice.

2.3 Feature Scaling

- Apply **standardization** or **normalization** to numerical features.
- Explain:
Why feature scaling is important in regression models

Step 3: Exploratory Data Analysis (EDA)

Perform EDA to understand the dataset:

1. Plot the distribution of the **charges** variable.
2. Analyze the relationship between:
 - o age vs charges
 - o bmi vs charges
 - o smoker vs charges
3. Generate a **correlation heatmap**.
4. Highlight **two key insights** obtained from the analysis.

Step 4: Model Building

4.1 Data Splitting

- Split the dataset into:
 - Training set (80%)
 - Testing set (20%)

4.2 Model Training

- Train a **Multiple Linear Regression** model.
- Display model coefficients and intercept.
- Write the regression equation

Step 5: Model Evaluation

Evaluate the model on the test dataset using:

- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- R-squared (R^2)

Explain what each metric indicates about model performance.

Task 2: Customer Term Deposit Prediction

You are provided with the **Bank Marketing Dataset**, which contains information collected from a Portuguese bank's direct marketing campaign.

<https://www.kaggle.com/datasets/janiobachmann/bank-marketing-dataset>

The goal is to **build a classification model** to predict whether a customer will **subscribe to a term deposit (yes / no)** based on socio-demographic and campaign features.

Data Loading & Inspection (3 Marks)

- Load the dataset into a Pandas DataFrame.
- Display the first five rows using `.head()`.
- Use `.info()` and `.describe()` to summarize the dataset.
- Identify:
 - The **target variable**
 - The **feature variables**

Data Preprocessing (4 Marks)

- Handle missing values (if any), and **justify your approach** (*e.g., mean/median/mode imputation or row removal*).
- Encode categorical variables such as:
 - job
 - marital
 - education
 - etc.using **One-Hot Encoding or Label Encoding**.

Feature Analysis & Handling Class Imbalance (3 Marks)

- Plot the **class distribution** of the target variable (y).
- If the dataset is imbalanced, apply **one method** to address class imbalance:
 - Oversampling
 - Undersampling
 - Class weights
- Briefly explain **why handling class imbalance may be necessary**.

Model Building (3 Marks)

- Split the dataset into:
 - Training set (80%)
 - Testing set (20%)
- Train **on classification model**:
 - Logistic Regression

Model Evaluation (2 Marks)

Evaluate the model using:

- **Accuracy**
- **Precision, Recall, F1-Score, Classification Matrix**

In one sentence, explain which evaluation metric is most suitable for this classification task and why.

Task 3:

- Discuss the following (Any 3)
- Overfitting
- Under fitting
- Bias
- Variance