Assignment-2 Lab-3 & 4

Atharva Date-B22AI045

Question 1

Task 1

In this task, the initial steps involve importing necessary libraries, loading the Titanic dataset, and performing basic exploratory data analysis.

- **Data Loading:**
  - The Titanic dataset is loaded from the provided URL using `pd.read_csv`.
- **Data Overview:**
  - The first 5 rows of the dataset are displayed using `df.head()` to get a quick overview.
  - `df.info()` is used to check for missing values and datatypes.
  - `df.isnull().sum()` is employed to identify the number of missing values in each column.
- **Grouping and Calculating Means:**
  - The mean of the 'Age' column is calculated based on the grouping of 'Pclass,' 'Sex,' and 'Survived' using `df.groupby(['Pclass','Sex','Survived'])['Age'].agg(['mean'])`.
  - Further refinement of the mean calculation is done by filling missing 'Age' values based on specific conditions.
- **Data Imputation:**
  - Missing values in the 'Age' column are filled based on the conditions - sex, Pclass and additionally survived. ***I included one more characteristic so that it would give more accuracy.***
- **Data Cleanup:**
  - The 'Cabin' column is dropped due to many missing values.
  - Rows with any remaining missing values are dropped using `df.dropna()`.
- **Encoding Categorical Variables:**
  - The 'Embarked' column is encoded numerically using the `encode_embarked()` function.
  - The 'Sex' column is replaced with numerical values (0 for female, 1 for male).

Types of data-

- **Ordinal Features:**
  - `Name`
  - `Passengerid`
- **Nominal Features:**
  - `Ticket`
- **Categorical Features:**
  - `Survived`: Represents whether a passenger survived (1) or did not survive (0).
  - `Sex`
  - `Pclass`
  - `SibSp`
  - `Embarked`
  - `Parch`

Task 2

This task involves the implementation of functions for entropy calculation and information gain.

- **Entropy Calculation:**
  - The `entropy()` function calculates the entropy of a given set of labels.
- **Information Gain Calculation:**
  - The `information_gain()` function calculates the information gain for a specific feature and threshold.

Task 3

In this task, a function `conTocat()` is implemented to find the best split for a decision tree.

- **Finding Best Split:**
  - The `conTocat()` function iterates through unique values of a feature and calculates information gain to find the best split.

I have not used this function exactly in the decision tree but I have used the concept of this function in it.

Task 4

Node Class:

- Represents a node in the decision tree.

DecisionTree Class:

- Represents the decision tree and contains methods for entropy calculation, growing the tree recursively, finding the best split, and training the tree using the `fit()` method.
- Predicting data by traversing the tree.

Task 5

Inference Function:

- The `infer()` function is implemented to predict the label for a single data point.

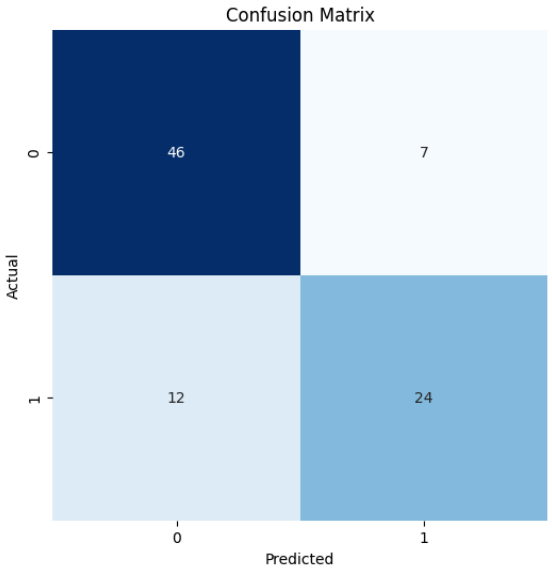
```
Test accuracy =
0.7865168539325843

Validation accuracy =
0.8033707865168539

Train accuracy =
0.9035369774919614
```

Task 6

The accuracy of the model was calculated on both the training and testing sets. And as expected the training set has high accuracy.

Task 7

`confusion_matrix` from `sklearn.metrics` is used to generate the confusion matrix. Confusion matrix was created using seaborn to visualize the performance of the model on the testing set which shows TP, TN, FP, FN.



Task 8

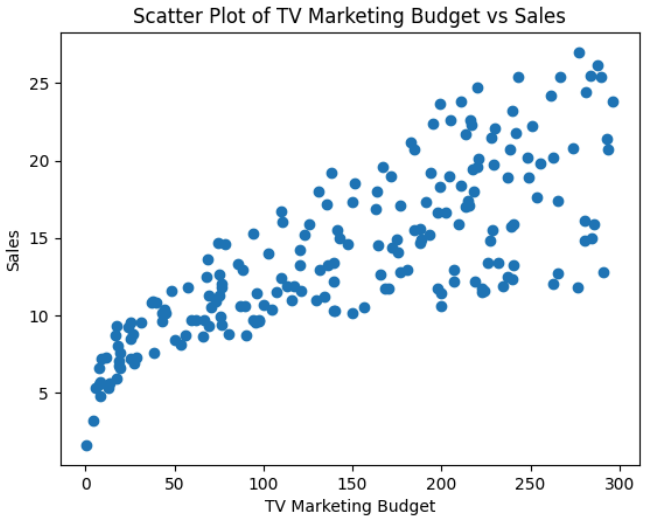
Precision, recall, and F1-score were calculated and printed for the model's performance on the testing set.

```
Precision: 0.7855
Recall: 0.7865
F1-score: 0.7834
```

Question 2

Task 1

For the second part, a dataset related to TV marketing and sales was loaded, and a scatter plot was created to visualize the relationship between TV marketing budget and sales. Basic statistics for both variables were calculated.



Data Overview:

- Mean and standard deviation for both 'TV' and 'Sales' columns are calculated.
- You can also use `describe()`

```
Mean TV Marketing Budget: 147.0425
Standard Deviation of TV Marketing Budget: 85.85423631490808
Mean Sales: 14.0225
Standard Deviation of Sales: 5.217456565710478
```

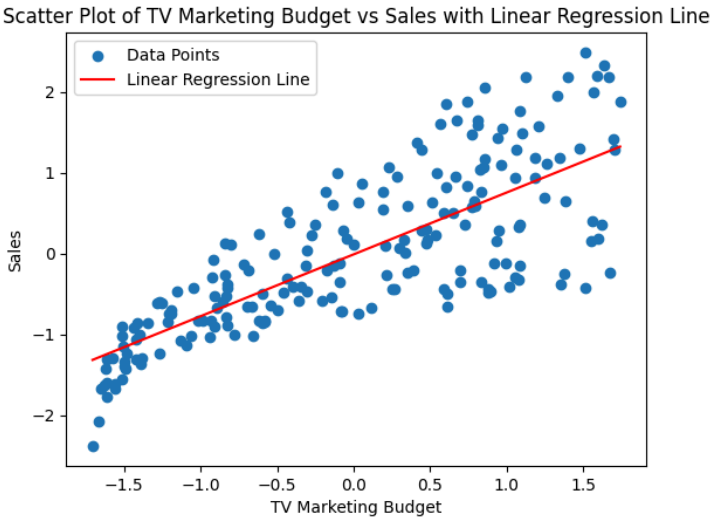
Maximum sale is 27, when the budget lies between 200 - 300.

Task 2

Handling Missing Values:

$$Z = \frac{(X - \mu)}{\sigma}$$

- Missing values are checked using `data.isnull().sum()`.
- The `z_transform` function is applied to standardize the 'TV' and 'Sales' columns.
- No null values were observed in the dataset.



Task 3

The gradient descent algorithm was implemented for univariate linear regression. The hypothesis function, cost function, and gradient descent function were defined. A graph with the line was

then plotted against the scatter plot of data. The final weight and bias is as follows.

```
Final Weights and bias : 0.7656555838856116 -0.011694050588820 67
```

Task 4

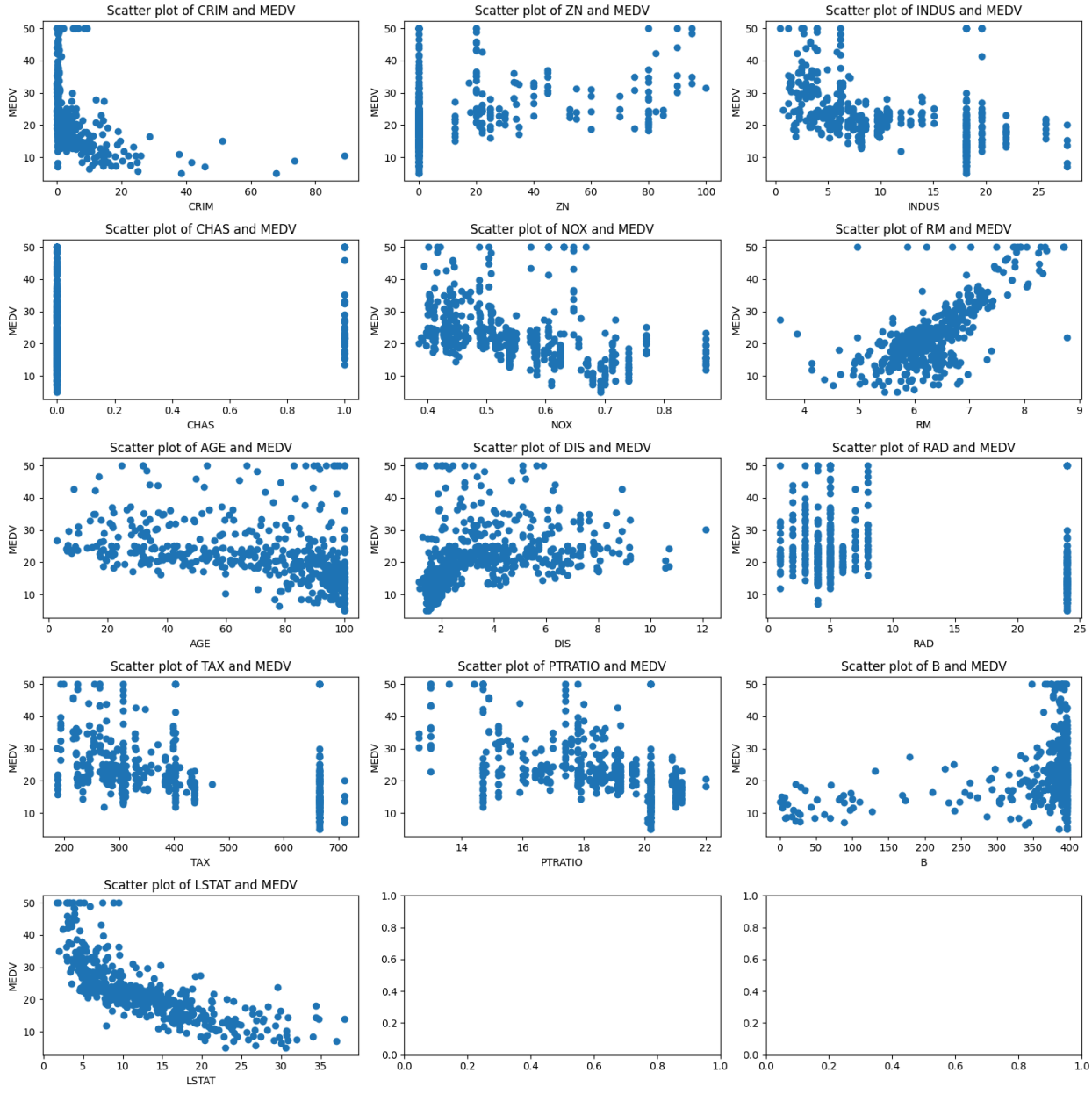
Predictions were made on the testing set, and mean squared error (MSE) and mean absolute error (MAE) were calculated..

```
Mean Squared Error: 0.3748699294331526
Mean Absolute Error: 0.46850797375410025
```

Question 3

Task 1

For the third part, the Boston Housing dataset was loaded, and scatter plots were created to visualize the relationship between each feature and the 'MEDV' variable.



```
Mean of CRIM is 3.613523557312254
STD of CRIM is 8.60154510533249
Mean of ZN is 11.363636363636363
STD of ZN is 23.32245299451514
Mean of INDUS is 11.13677865612648
STD of INDUS is 6.860352940897585
Mean of CHAS is 0.0691699604743083
STD of CHAS is 0.25399404134041037
Mean of NOX is 0.5546950592885376
STD of NOX is 0.11587767566755595
Mean of RM is 6.284634387351779
STD of RM is 0.7026171434153233
Mean of AGE is 68.57490118577076
STD of AGE is 28.148861406903617
Mean of DIS is 3.795042687747036
STD of DIS is 2.105710126627611
Mean of RAD is 9.549407114624506
STD of RAD is 8.707259384239366
Mean of TAX is 408.2371541501976
STD of TAX is 168.53711605495903
Mean of PTRATIO is 18.455533596837945
STD of PTRATIO is 2.1649455237144406
Mean of B is 356.6740316205534
STD of B is 91.29486438415783
Mean of LSTAT is 12.653063241106722
STD of LSTAT is 7.141061511348571
Mean of MEDV is 22.532806324110677
STD of MEDV is 9.197104087379818
```

The above are the mean and standard deviation of all classes.Which were too calculated

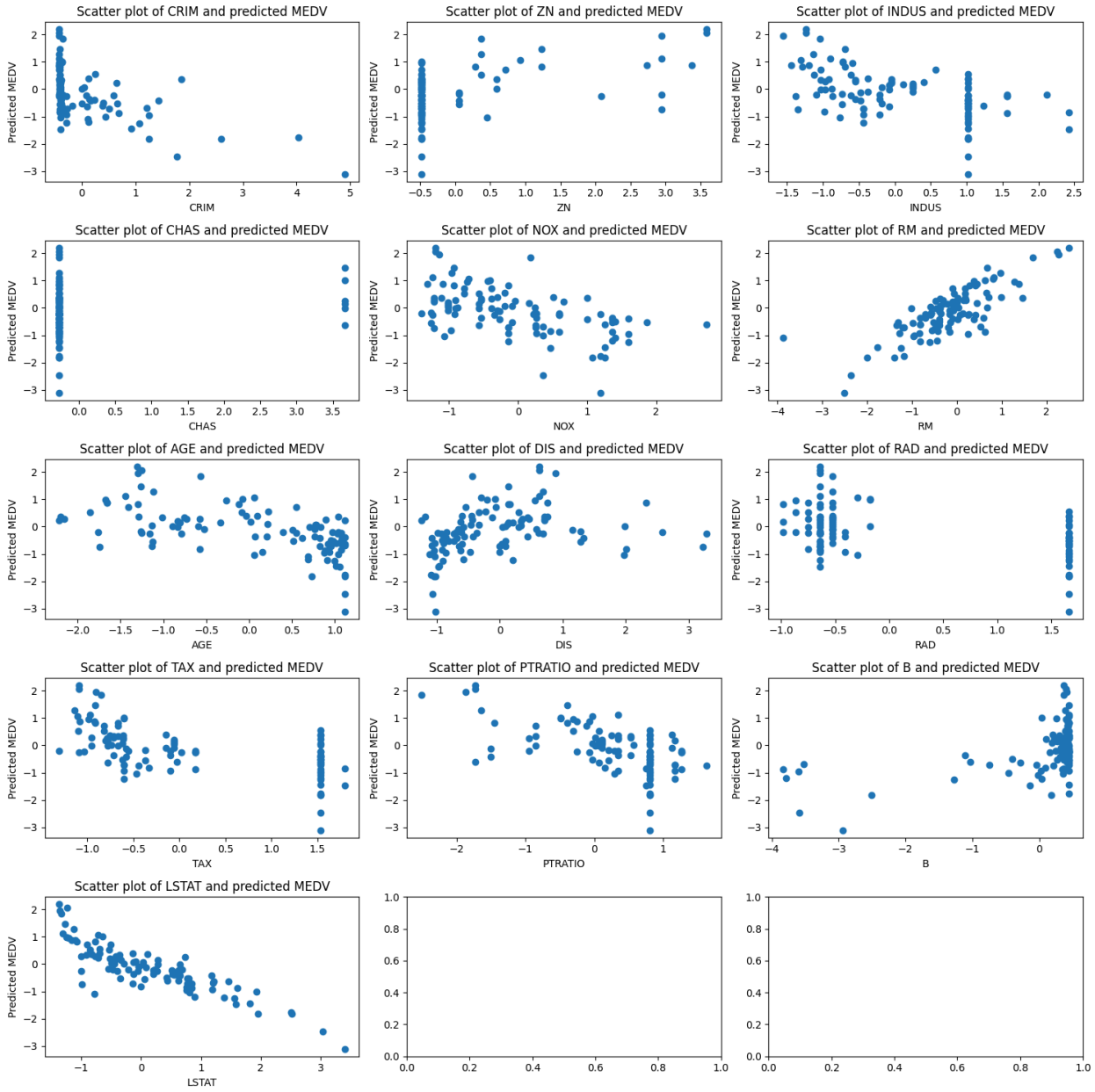
I have also used `describe()`
It is also observed the data is capped to some extent looking at the graph of CRIM. Thus it may result in loss in precision.

Task 2

Basic statistics for each feature were calculated, and z-score normalization was applied to the dataset and splitting was done for the data. No null values are present in the dataset

Task 3

The gradient descent algorithm was applied to perform multivariate linear regression.A scatter plot is displays that contains predicted values of corresponding X_ test-



Task 4

Predictions were made on the testing set, and MSE and MAE were calculated

```
Mean Squared Error: 0.28720068788225034
Mean Absolute Error: 0.346763842440638
```