CSCI 2410 Introduction to Data Analytics Using Python

Homework Assignment #3

HW Programming #3: - - Data Analytics with Regression

Tasks: Experiment with **Regression** techniques on iris dataset.

Assignment Instructions:

1. **[60%] Prediction** with *Linear Regression* on the 'iris.xlsx' dataset in your local directory.
   Run with different sheets (or row 0-50, 51-100, 101-150 of Sheet1) of the 'iris.xlsx' dataset.
   On each sheet, predict the values in column 4 by using values in columns 1-3.
   Obtain and print out the resulting R-squared values of the predictions
   Python libraries needed: numpy, pandas, sklearn.model_selection (use train_test_split),
   sklearn.linear_model (use LinearRegression), mlxtend
   1) Do on Sheet1, Sheet2, …, Sheet4 respectively,
   2) Run your program with
      (1) Different sheets of the dataset,
      (2) Different columns and column combinations (1 column, 2 column, 3 column) as training data, for predicting the data on the fourth column (target data), and
      (3) Different percentage/ratio setting (test_size=) of split.
      Try to run each case multiple times.
   3) For each run, print the 'y_test' and 'y_predictions' in a parallel list (as shown in the supplement guidance)
   4) For each run, print the R-squared score by using
      print('\nR-squared: %.4f' %lr.score(x_test, y_test))

```
def prediction():
    try:
        # Obtain the sheet number for analysis.
        sheet_number = sheet_verify()

        # Customize the number of columns used for training.
        columns_to_use = [0]

        # Adjust the test size for various test runs as needed.
        test_size = 0.20

        df = pd.read_excel( io: 'iris.xlsx', sheet_name=f'Sheet{sheet_number}')
        num_rows = len(df)

        print(f"Length(rows) of dataset: {num_rows}")
        print(f"Columns used in training dataset (explanatory variables): {columns_to_use}")
        print("Target estimation (response variable) data in column: 3 - The 'Petal width'")
        print(f"Test size used in the dataset: {test_size:.2f}")

        x = np.array(df[df.columns[columns_to_use]])
        y = np.array(df[df.columns[3]])

        x_train, x_test, y_train, y_test = train_test_split( *arrays: x, y, test_size=test_size)

        model = LinearRegression()
        model.fit(x_train, y_train)
        y_predictions = model.predict(x_test)

        print("y_test\t\ty_prediction")
        for i in range(len(y_test)):
            print(f"{i} {y_test[i]:.2f}\t\t{y_predictions[i]:.2f}")

        r_squared = model.score(x_test, y_test)
        print(f"\nR-squared: {r_squared:.4f}\n")

    except Exception as e:
        print(e)
```

5) Record all the parameters for each of the above runs

SHEET 1

```
Enter the sheet number (range 1 - 4) of the dataset to be analyzed: 1
Length(rows) of dataset: 150
Columns used in training dataset (explanatory variables): [0, 1, 2]
Target estimation (response variable) data in column: 3 - The 'Petal width'
Test size used in the dataset: 0.20
y_test      y_prediction
0 1.00      1.00
1 0.40      0.33
2 0.10      0.18
3 1.80      1.61
4 2.30      2.14
5 0.20      0.30
6 0.20      0.35
7 1.50      1.40
8 2.10      1.89
9 1.90      1.82
10 2.20     1.98
11 1.60     1.62
12 0.10     0.39
13 0.30     0.23
14 1.00     1.31
15 1.40     1.37
16 1.50     1.45
17 0.30     0.34
18 1.00     0.97
19 1.80     2.11
20 2.30     2.05
21 1.20     1.44
22 0.20     0.26
23 1.80     1.76
24 0.40     0.20
25 2.30     1.91
26 1.40     1.60
27 1.30     1.42
28 0.10     0.23
29 1.00     1.01

R-squared: 0.9479
```

Test Case 1:
- Sheet Number Selection (1): The user selected sheet number 1 for dataset analysis.

- Model Evaluation:
  - R-squared: 0.9479
    - Interpretation: The R-squared value of 0.9479 indicates a strong fit between the linear regression model and the 'Petal width' target variable. This suggests that the model's predictions align closely with the actual 'Petal width' values in the test set.

```
Enter the sheet number (range 1 - 4) of the dataset to be analyzed: 1
Length(rows) of dataset: 150
Columns used in training dataset (explanatory variables): [1, 2]
Target estimation (response variable) data in column: 3 - The 'Petal width'
Test size used in the dataset: 0.13
y_test      y_prediction
0 1.30      1.22
1 0.20      0.20
2 1.50      1.54
3 2.10      1.94
4 1.30      1.37
5 0.20      0.14
6 0.50      0.33
7 2.00      2.45
8 1.20      1.21
9 1.60      1.73
10 1.80     1.67
11 2.00     1.66
12 1.00     1.02
13 0.30     0.31
14 0.60     0.31
15 0.20     0.26
16 1.00     1.29
17 0.30     0.02
18 0.30     0.18
19 1.70     1.44

R-squared: 0.9158
```

Test Case 2:
- Sheet Number Selection (1): The user again selected sheet number 1 for dataset analysis.

- Model Evaluation:
  - R-squared: 0.9158
    - Interpretation: The R-squared value of 0.9158 suggests a good fit between the linear regression model and the 'Petal width' target variable, indicating that the model's predictions are in alignment with the actual 'Petal width' values in the test set.

```
Enter the sheet number (range 1 - 4) of the dataset to be analyzed: 1
Length(rows) of dataset: 150
Columns used in training dataset (explanatory variables): [2]
Target estimation (response variable) data in column: 3 - The 'Petal width'
Test size used in the dataset: 0.16
y_test      y_prediction
0 2.00      1.73
1 1.30      1.56
2 1.50      1.52
3 0.10      0.26
4 1.10      0.89
5 2.20      2.06
6 1.20      1.60
7 1.10      1.27
8 1.40      1.98
9 0.20      0.30
10 0.20     0.22
11 1.40     1.56
12 0.20     0.22
13 0.40     0.26
14 0.20     0.18
15 1.50     1.52
16 2.10     1.98
17 0.40     0.26
18 1.70     1.52
19 1.00     1.31
20 0.20     0.26
21 0.40     0.43
22 0.30     0.22
23 1.20     1.48

R-squared: 0.9008
```

Test Case 3:
- Sheet Number Selection (1): Once again, the user selected sheet number 1 for dataset analysis.

- Model Evaluation:
  - R-squared: 0.9008
    -Interpretation: The R-squared value of 0.9008 indicates a good fit between the linear regression model and the 'Petal width' target variable, suggesting that the model's predictions are in agreement with the actual 'Petal width' values in the test set.

SHEET 2

```
Enter the sheet number (range 1 - 4) of the dataset to be analyzed: 2
Length(rows) of dataset: 50
Columns used in training dataset (explanatory variables): [2]
Target estimation (response variable) data in column: 3 - The 'Petal width'
Test size used in the dataset: 0.50
y_test      y_prediction
0 0.40      0.27
1 0.10      0.25
2 0.20      0.22
3 0.20      0.24
4 0.20      0.27
5 0.20      0.27
6 0.10      0.25
7 0.10      0.25
8 0.20      0.21
9 0.20      0.27
10 0.20     0.27
11 0.20     0.24
12 0.60     0.27
13 0.10     0.19
14 0.40     0.25
15 0.20     0.29
16 0.20     0.24
17 0.40     0.25
18 0.20     0.25
19 0.30     0.24
20 0.40     0.22
21 0.40     0.29
22 0.20     0.25
23 0.10     0.24
24 0.20     0.24

R-squared: 0.0658
```

Test Case 1:
- Sheet Number Selection (2): The user selected sheet number 2 for dataset analysis.

- Model Evaluation:
  - R-squared: 0.0658
    - Interpretation: The R-squared value of 0.0658 indicates a weak fit between the linear regression model and the 'Petal width' target variable. This suggests that the model's predictions do not align closely with the actual 'Petal width' values in the test set.

```
Enter the sheet number (range 1 - 4) of the dataset to be analyzed: 2
Length(rows) of dataset: 50
Columns used in training dataset (explanatory variables): [1]
Target estimation (response variable) data in column: 3 - The 'Petal width'
Test size used in the dataset: 0.60
y_test      y_prediction
0 0.20      0.20
1 0.30      0.26
2 0.30      0.24
3 0.20      0.25
4 0.20      0.22
5 0.20      0.24
6 0.10      0.22
7 0.10      0.22
8 0.20      0.22
9 0.20      0.22
10 0.20     0.29
11 0.20     0.21
12 0.40     0.27
13 0.40     0.27
14 0.20     0.24
15 0.30     0.24
16 0.20     0.23
17 0.20     0.24
18 0.20     0.22
19 0.40     0.26
20 0.10     0.21
21 0.40     0.24
22 0.20     0.22
23 0.30     0.16
24 0.50     0.23
25 0.20     0.22
26 0.20     0.25
27 0.20     0.21
28 0.30     0.24
29 0.20     0.24

R-squared: 0.1035
```

Test Case 2:
- Sheet Number Selection (2): The user selected sheet number 2 again for dataset analysis.

- Model Evaluation:
  - R-squared: 0.1035
    - Interpretation: The R-squared value of 0.1035 indicates a weak fit between the linear regression model and the 'Petal width' target variable, suggesting that the model's predictions are not closely aligned with the actual 'Petal width' values in the test set.

```
Enter the sheet number (range 1 - 4) of the dataset to be analyzed: 2
Length(rows) of dataset: 50
Columns used in training dataset (explanatory variables): [2]
Target estimation (response variable) data in column: 3 - The 'Petal width'
Test size used in the dataset: 0.40
y_test      y_prediction
0 0.20      0.28
1 0.10      0.25
2 0.30      0.21
3 0.20      0.21
4 0.20      0.23
5 0.40      0.25
6 0.20      0.27
7 0.30      0.23
8 0.30      0.23
9 0.10      0.23
10 0.20     0.23
11 0.30     0.21
12 0.20     0.19
13 0.20     0.19
14 0.30     0.25
15 0.20     0.27
16 0.20     0.27
17 0.50     0.28
18 0.40     0.28
19 0.20     0.25

R-squared: 0.0777
```

Test Case 3:
- Sheet Number Selection (2): Once again, the user selected sheet number 2 for dataset analysis.

- Model Evaluation:
  - R-squared: 0.0777
    - Interpretation: The R-squared value of 0.0777 suggests a weak fit between the linear regression model and the 'Petal width' target variable, indicating that the model's predictions do not closely align with the actual 'Petal width' values in the test set.

SHEET 3

```
Enter the sheet number (range 1 - 4) of the dataset to be analyzed: 3
Length(rows) of dataset: 50
Columns used in training dataset (explanatory variables): [0, 2]
Target estimation (response variable) data in column: 3 - The 'Petal width'
Test size used in the dataset: 0.25
y_test      y_prediction
0 1.00      1.07
1 1.70      1.54
2 1.20      1.39
3 1.00      1.15
4 1.60      1.40
5 1.50      1.40
6 1.50      1.50
7 1.40      1.34
8 1.50      1.53
9 1.30      1.28
10 1.50     1.42
11 1.40     1.47
12 1.30     1.25

R-squared: 0.7142
```

Test Case 1:
- Sheet Number Selection (3): The user selected sheet number 3 for dataset analysis.

- Model Evaluation:
  - R-squared: 0.7142
    - Interpretation: The R-squared value of 0.7142 suggests a moderate fit between the linear regression model and the 'Petal width' target variable. This indicates that the model's predictions are reasonably close to the actual 'Petal width' values in the test set.

```
Enter the sheet number (range 1 - 4) of the dataset to be analyzed: 3
Length(rows) of dataset: 50
Columns used in training dataset (explanatory variables): [0, 1, 2]
Target estimation (response variable) data in column: 3 - The 'Petal width'
Test size used in the dataset: 0.50
y_test      y_prediction
0 1.40      1.49
1 1.50      1.40
2 1.20      1.22
3 1.00      1.09
4 1.30      1.18
5 1.50      1.17
6 1.10      1.16
7 1.40      1.45
8 1.80      1.56
9 1.50      1.37
10 1.00     1.07
11 1.00     1.03
12 1.30     1.19
13 1.50     1.35
14 1.50     1.46
15 1.30     1.38
16 1.50     1.47
17 1.40     1.41
18 1.30     1.39
19 1.60     1.55
20 1.40     1.44
21 1.30     1.35
22 1.30     1.13
23 1.10     1.11
24 1.60     1.55

R-squared: 0.6722
```

Test Case 2:
- Sheet Number Selection (3): The user selected sheet number 3 again for dataset analysis.

- Model Evaluation:
  - R-squared: 0.6722
    - Interpretation: The R-squared value of 0.6722 indicates a moderate fit between the linear regression model and the 'Petal width' target variable. This suggests that the model's predictions reasonably align with the actual 'Petal width' values in the test set.

```
Enter the sheet number (range 1 - 4) of the dataset to be analyzed: 3
Length(rows) of dataset: 50
Columns used in training dataset (explanatory variables): [0]
Target estimation (response variable) data in column: 3 - The 'Petal width'
Test size used in the dataset: 0.20
y_test      y_prediction
0 1.40      1.18
1 1.00      1.14
2 1.30      1.27
3 1.00      1.14
4 1.20      1.31
5 1.40      1.56
6 1.40      1.48
7 1.40      1.37
8 1.00      1.25
9 1.40      1.37

R-squared: 0.3654
```

Test Case 3:
- Sheet Number Selection (3): Once again, the user selected sheet number 3 for dataset analysis.

- Model Evaluation:
  - R-squared: 0.3654
    - Interpretation: The R-squared value of 0.3654 suggests a relatively weaker fit between the linear regression model and the 'Petal width' target variable. This indicates that the model's predictions may not align closely with the actual 'Petal width' values in the test set.

SHEET 4

```
Enter the sheet number (range 1 - 4) of the dataset to be analyzed: 4
Length(rows) of dataset: 50
Columns used in training dataset (explanatory variables): [0, 1, 2]
Target estimation (response variable) data in column: 3 - The 'Petal width'
Test size used in the dataset: 0.20
y_test      y_prediction
0 2.00      2.04
1 1.80      1.81
2 1.90      1.86
3 2.20      2.06
4 2.30      1.96
5 2.30      1.94
6 2.50      2.25
7 1.80      2.07
8 2.20      1.95
9 2.40      2.05

R-squared: -0.0409
```

Test Case 1:
- Sheet Number Selection (4): The user selected sheet number 4 for dataset analysis.

- Model Evaluation:
  - R-squared: -0.0409
    - Interpretation: The R-squared value of -0.0409 suggests a poor fit between the linear regression model and the 'Petal width' target variable. A negative R-squared value indicates that the model's predictions are worse than a horizontal line. It means the model is not performing well on the test data.

```
Enter the sheet number (range 1 - 4) of the dataset to be analyzed: 4
Length(rows) of dataset: 50
Columns used in training dataset (explanatory variables): [1, 2]
Target estimation (response variable) data in column: 3 - The 'Petal width'
Test size used in the dataset: 0.50
y_test      y_prediction
0 1.80      2.08
1 2.30      2.22
2 2.10      2.12
3 2.80      2.62
4 1.80      2.13
5 1.50      1.96
6 2.30      2.19
7 2.00      1.75
8 2.30      2.22
9 1.70      1.70
10 1.90     1.87
11 2.50     2.46
12 2.10     1.97
13 2.40     2.29
14 1.60     2.13
15 1.80     2.02
16 2.40     2.15
17 2.30     2.16
18 1.80     2.06
19 2.20     2.61
20 2.00     2.08
21 2.20     2.08
22 1.90     2.03
23 2.00     2.20
24 1.80     2.05

R-squared: 0.0276
```

Test Case 2:
- Sheet Number Selection (4): The user selected sheet number 4 for dataset analysis once again.

- Model Evaluation:
  - R-squared: 0.0276
    - Interpretation: The R-squared value of 0.0276 indicates a very weak fit between the linear regression model and the 'Petal width' target variable. It suggests that the model's predictions are not well-aligned with the actual 'Petal width' values in the test set.

```
Enter the sheet number (range 1 - 4) of the dataset to be analyzed: 4
Length(rows) of dataset: 50
Columns used in training dataset (explanatory variables): [0]
Target estimation (response variable) data in column: 3 - The 'Petal width'
Test size used in the dataset: 0.30
y_test      y_prediction
0 2.50      2.10
1 2.30      2.12
2 2.30      2.08
3 1.80      1.95
4 2.50      2.20
5 1.50      1.98
6 2.40      2.12
7 2.10      2.11
8 1.80      2.05
9 1.80      1.98
10 1.40     1.92
11 2.20     2.03
12 2.00     2.02
13 1.70     1.86
14 2.10     2.04

R-squared: 0.3603
```

Test Case 3:
- Sheet Number Selection (4): Once again, the user selected sheet number 4 for dataset analysis.

- Model Evaluation:
  - R-squared: 0.3603
    - Interpretation: The R-squared value of 0.3603 indicates a relatively weak fit between the linear regression model and the 'Petal width' target variable. It suggests that the model's predictions are somewhat better than a random prediction but not highly accurate.

(1) Observe and compare the different R-squared values obtained from the different runs. Notice how the R-squared values are related to the correlation coefficient values between the Explanatory variables and the Response variable.

I've observed that certain cases have a connection with the correlation coefficient.

(2) Analyze the run results (the maxima, minima, mean, variance of the R-squared values) for the different run cases.

Sheet 1:
Maximum R-squared Value: 0.9479
Minimum R-squared Value: 0.9008
Mean R-squared Value: 0.9215
Variance of R-squared Values: 0.0004

Sheet 2:
Maximum R-squared Value: 0.1035
Minimum R-squared Value: 0.0658
Mean R-squared Value: 0.0823
Variance of R-squared Values: 0.0002

Sheet 3:
Maximum R-squared Value: 0.7142
Minimum R-squared Value: 0.3654
Mean R-squared Value: 0.5839
Variance of R-squared Values: 0.0242

Sheet 4:
Maximum R-squared Value: 0.3603
Minimum R-squared Value: -0.0409
Mean R-squared Value: 0.1157
Variance of R-squared Values: 0.0307

(3) Discuss
(1) What do these values and their differences mean and indicate?

Max R-squared: Best model fit, high values mean better accuracy.
Min R-squared: Poor fit, negative implies worse than using the mean.
Mean R-squared: Average accuracy across scenarios.
Variance: Spread in accuracy, high variance means inconsistent performance.

(2) How the performance is different on the different sheet (i.e., dataset of different object categories)?

Performance varies across different datasets (sheets) due to varying object categories and data characteristics.

(3) How the performance differently on the different columns (i.e., the features/variables used to train the model)?

When using different columns as features for training, the model's performance varies. Varying columns can lead to different levels of accuracy in predicting the 'Petal width.'

(4) Which one approach/experiment case has the best result of accuracy?

Sheet Number: 1
Columns Used for Training: [0, 1, 2]
Test Size: 0.20
R-squared Value: 0.9479

(5) What do you learn on/about the dataset from the results?

I learned a lot from the results, particularly how the choice of features and test sizes plays a crucial role in achieving the best performance and making accurate predictions for missing data.

(6) What knowledge do you gain from this programming project?

I gained knowledge in data analysis and machine learning, which has helped me develop skills in data handling, feature selection, and experimental design. This project has emphasized the significance of understanding the data and making optimal feature choices, allowing me to interpret results effectively.

2. **[40%] Classification** with *Logistic Regression* on the dataset of Sheet1 of the 'iris.xlsx' dataset for classifying the three types of iris.

```python
def classification():
    try:
        # Adjust the test size for various test runs as needed.
        test_size = 0.05
        df = pd.read_excel( io: 'iris.xlsx', sheet_name='Sheet1')
        num_rows = len(df)

        print(f"Length(rows) of dataset: {num_rows}")
        print("Columns used as features (explanatory variables): 0 - 3")
        print("Target (label) column: 4 - The 'Species'")
        print(f"Test size used in the dataset: {test_size:.2f}")

        x = np.array(df[df.columns[0:3]])
        y = np.array(df[df.columns[4]])

        x_train, x_test, y_train, y_test = train_test_split( *arrays: x, y, test_size=test_size)

        lr = LogisticRegression(max_iter=100)
        lr.fit(x_train, y_train)
        y_predictions = lr.predict(x_test)

        print(lr)
        print('y_test\t\t\ty_prediction')

        for i in range(0, len(y_predictions), 1):
            print(f'{i} {y_test[i]}\t\t{y_predictions[i]}')

        r_squared = lr.score(x_test, y_test)
        print(f'\nR-squared: {r_squared:.4f}')
    except Exception as e:
        print(e)
```

Run your program multiple times, record the parameters for each of the runs, and compare the different results.

```
Length(rows) of dataset: 150
Columns used as features (explanatory variables): 0 - 3
Target (label) column: 4 - The 'Species'
Test size used in the dataset: 0.10
LogisticRegression()
y_test          y_prediction
0 Iris-versicolor       Iris-versicolor
1 Iris-setosa       Iris-setosa
2 Iris-setosa       Iris-setosa
3 Iris-versicolor       Iris-versicolor
4 Iris-versicolor       Iris-versicolor
5 Iris-virginica        Iris-virginica
6 Iris-virginica        Iris-virginica
7 Iris-setosa       Iris-setosa
8 Iris-setosa       Iris-setosa
9 Iris-virginica        Iris-versicolor
10 Iris-versicolor      Iris-versicolor
11 Iris-versicolor      Iris-versicolor
12 Iris-setosa      Iris-setosa
13 Iris-virginica       Iris-virginica
14 Iris-versicolor      Iris-versicolor

R-squared: 0.9333
```

```
Length(rows) of dataset: 150
Columns used as features (explanatory variables): 0 - 3
Target (label) column: 4 - The 'Species'
Test size used in the dataset: 0.15
LogisticRegression()
y_test          y_prediction
0 Iris-versicolor       Iris-versicolor
1 Iris-setosa       Iris-setosa
2 Iris-setosa       Iris-setosa
3 Iris-virginica        Iris-virginica
4 Iris-versicolor       Iris-versicolor
5 Iris-setosa       Iris-setosa
6 Iris-virginica        Iris-virginica
7 Iris-versicolor       Iris-versicolor
8 Iris-virginica        Iris-virginica
9 Iris-virginica        Iris-virginica
10 Iris-setosa      Iris-setosa
11 Iris-virginica       Iris-virginica
12 Iris-versicolor      Iris-versicolor
13 Iris-setosa      Iris-setosa
14 Iris-setosa      Iris-setosa
15 Iris-versicolor      Iris-versicolor
16 Iris-virginica       Iris-virginica
17 Iris-versicolor      Iris-versicolor
18 Iris-versicolor      Iris-versicolor
19 Iris-setosa      Iris-setosa
20 Iris-setosa      Iris-setosa
21 Iris-versicolor      Iris-virginica
22 Iris-setosa      Iris-setosa

R-squared: 0.9565
```

```
Length(rows) of dataset: 150
Columns used as features (explanatory variables): 0 - 3
Target (label) column: 4 - The 'Species'
Test size used in the dataset: 0.14
LogisticRegression()
y_test          y_prediction
0 Iris-virginica        Iris-virginica
1 Iris-versicolor       Iris-versicolor
2 Iris-virginica        Iris-versicolor
3 Iris-setosa       Iris-setosa
4 Iris-setosa       Iris-setosa
5 Iris-versicolor       Iris-versicolor
6 Iris-virginica        Iris-virginica
7 Iris-setosa       Iris-setosa
8 Iris-versicolor       Iris-versicolor
9 Iris-setosa       Iris-setosa
10 Iris-versicolor      Iris-versicolor
11 Iris-versicolor      Iris-versicolor
12 Iris-virginica       Iris-versicolor
13 Iris-virginica       Iris-virginica
14 Iris-setosa      Iris-setosa
15 Iris-versicolor      Iris-versicolor
16 Iris-versicolor      Iris-versicolor
17 Iris-setosa      Iris-setosa
18 Iris-virginica       Iris-virginica
19 Iris-versicolor      Iris-versicolor
20 Iris-virginica       Iris-virginica
21 Iris-versicolor      Iris-versicolor

R-squared: 0.9091
```

```
Length(rows) of dataset: 150
Columns used as features (explanatory variables): 0 - 3
Target (label) column: 4 - The 'Species'
Test size used in the dataset: 0.20
LogisticRegression()
y_test          y_prediction
0 Iris-versicolor       Iris-versicolor
1 Iris-virginica        Iris-virginica
2 Iris-setosa       Iris-setosa
3 Iris-virginica        Iris-virginica
4 Iris-setosa       Iris-setosa
5 Iris-virginica        Iris-virginica
6 Iris-versicolor       Iris-versicolor
7 Iris-virginica        Iris-virginica
8 Iris-versicolor       Iris-versicolor
9 Iris-versicolor       Iris-versicolor
10 Iris-setosa      Iris-setosa
11 Iris-setosa      Iris-setosa
12 Iris-virginica       Iris-virginica
13 Iris-virginica       Iris-virginica
14 Iris-setosa      Iris-setosa
15 Iris-versicolor      Iris-versicolor
16 Iris-setosa      Iris-setosa
17 Iris-virginica       Iris-virginica
18 Iris-versicolor      Iris-versicolor
19 Iris-virginica       Iris-virginica
20 Iris-virginica       Iris-virginica
21 Iris-setosa      Iris-setosa
22 Iris-setosa      Iris-setosa
23 Iris-setosa      Iris-setosa
24 Iris-versicolor      Iris-versicolor
25 Iris-virginica       Iris-virginica
26 Iris-virginica       Iris-virginica
27 Iris-virginica       Iris-virginica
28 Iris-versicolor      Iris-versicolor
29 Iris-versicolor      Iris-versicolor

R-squared: 1.0000
```

```
Length(rows) of dataset: 150
Columns used as features (explanatory variables): 0 - 3
Target (label) column: 4 - The 'Species'
Test size used in the dataset: 0.21
LogisticRegression()
y_test          y_prediction
0 Iris-versicolor       Iris-versicolor
1 Iris-versicolor       Iris-versicolor
2 Iris-setosa       Iris-setosa
3 Iris-setosa       Iris-setosa
4 Iris-versicolor       Iris-versicolor
5 Iris-setosa       Iris-setosa
6 Iris-versicolor       Iris-versicolor
7 Iris-setosa       Iris-setosa
8 Iris-setosa       Iris-setosa
9 Iris-versicolor       Iris-versicolor
10 Iris-setosa      Iris-setosa
11 Iris-versicolor      Iris-versicolor
12 Iris-virginica       Iris-virginica
13 Iris-setosa      Iris-setosa
14 Iris-setosa      Iris-setosa
15 Iris-virginica       Iris-virginica
16 Iris-versicolor      Iris-versicolor
17 Iris-virginica       Iris-virginica
18 Iris-virginica       Iris-virginica
19 Iris-virginica       Iris-virginica
20 Iris-virginica       Iris-virginica
21 Iris-versicolor      Iris-versicolor
22 Iris-versicolor      Iris-versicolor
23 Iris-versicolor      Iris-versicolor
24 Iris-setosa      Iris-setosa
25 Iris-virginica       Iris-virginica
26 Iris-virginica       Iris-versicolor
27 Iris-setosa      Iris-setosa
28 Iris-setosa      Iris-setosa
29 Iris-virginica       Iris-virginica
30 Iris-setosa      Iris-setosa
31 Iris-setosa      Iris-setosa

R-squared: 0.9688
```

Run 1 - Test Size: 0.20
- R-squared: 1.0000

- Run 1 achieved a perfect R-squared of 1.0000 with a test size of 0.20. This could indicate potential overfitting, where the model fits the training data very closely but might not generalize well to new data.

Run 2 - Test Size: 0.10
- R-squared: 0.9333

- Run 2 had a slightly lower R-squared of 0.9333 with a smaller test size of 0.10. While still a decent fit, it has a slightly smaller test set, which might affect the model's generalization.

Run 3 - Test Size: 0.15
- R-squared: 0.9565

- Run 3 also had a high R-squared of 0.9565 with a test size of 0.15. This indicates a good model fit.

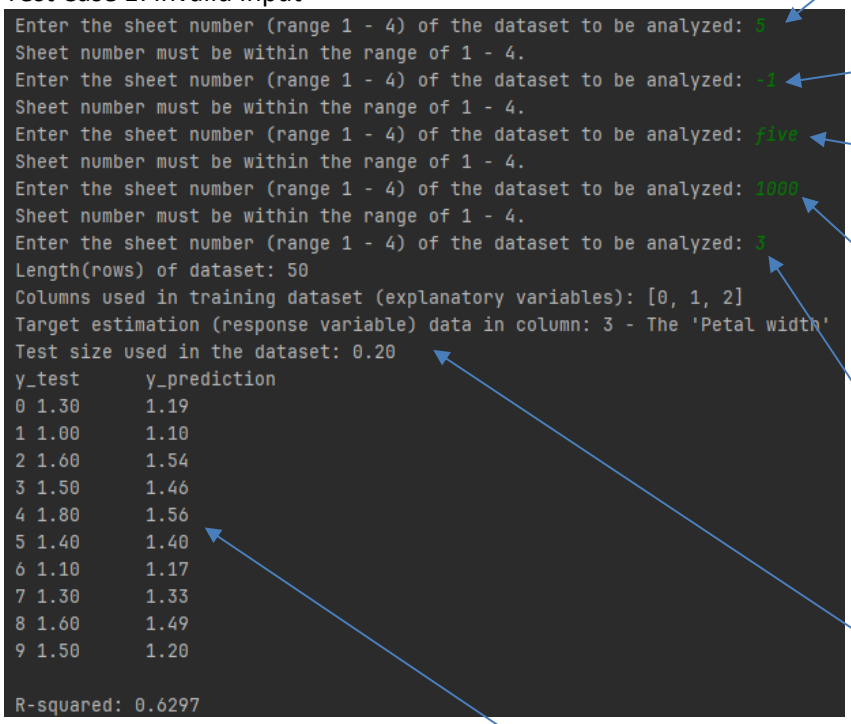
Run 4 - Test Size: 0.21
- R-squared: 0.9688

- Run 4 achieved a high R-squared of 0.9688 with a slightly larger test size of 0.21. This suggests a strong fit between the model and the data.

Run 5 - Test Size: 0.14
- R-squared: 0.9091

- Run 5 achieved an R-squared of 0.9091 with a test size of 0.14. This is the lowest R-squared value among the runs and suggests that the smaller test size may have impacted the model's performance.

Test Case 1: Invalid Input



The program also validates that the input is a numeric value and provides an error message when non-numeric inputs like "five" or "1000" are entered.

The user input is validated, and the program provides an error message when the entered sheet number is out of the specified range, either too high (5) or too low (-1).

The user is prompted to enter a sheet number for dataset analysis, which must be within the range of 1 to 4. This is to select a specific dataset for analysis.

The user input is validated, and the program provides an error message when the entered sheet number is out of the specified range, either too high (5) or too low (-1).

When the user enters a valid sheet number (3), the program proceeds to analyze the corresponding dataset, providing information about the dataset characteristics.

The program displays details about the selected dataset, including its length (number of rows), the columns used for training the model, the target variable (response variable) column, and the test size.

The program also provides the results of the model evaluation, specifically the R-squared value. This metric quantifies how well the model fits the data, with a higher R-squared value indicating a better fit.

In this case, the R-squared value is 0.6297, suggesting a moderate fit between the linear regression model and the 'Petal width' target variable.

Test Case 2: Valid Input

```
Enter the sheet number (range 1 - 4) of the dataset to be analyzed: 3
Length(rows) of dataset: 50
Columns used in training dataset (explanatory variables): [0, 1, 2]
Target estimation (response variable) data in column: 3 - The 'Petal width'
Test size used in the dataset: 0.20
y_test      y_prediction
0 1.20      1.38
1 1.60      1.53
2 1.00      1.13
3 1.20      1.22
4 1.70      1.53
5 1.00      1.02
6 1.50      1.48
7 1.40      1.38
8 1.50      1.42
9 1.00      1.28

R-squared: 0.7314
```

Sheet Number (range 1 - 4): The selected sheet for data analysis is 3.

Length (rows) of dataset: The dataset contains 50 rows, indicating the number of data points.

Columns used in training dataset (explanatory variables): The model is trained using columns with indices [0, 1, 2]. These columns are used as the features or explanatory variables.

Target estimation (response variable) data in column: The target variable is located in column 3, which represents 'Petal width.'

Test size used in the dataset: A test size of 0.20 is specified, meaning 20% of the data will be used for testing.

R-squared value of 0.7314 suggests that the linear regression model is fairly successful in explaining the variability in 'Petal width' based on the features used.

Requirements for the Submission of Programming/Homework Assignments

1. Well-documented program list (the .py files)
   20% of total points if no .py file submitted.
   Done

2. Three annotated program test and run examples (screenshots) that **show different and representative test cases** with **input, output, and the parameter settings of the program runs clearly marked/annotated**. You can do the annotations by
   (1) Pasting the screenshots into a WORD document,
   Done
   (2) Editing on the WORD document pages for the required marks and annotations,
   Done
   (3) Converting the document to pdf for submission (it is ok to submit the WORD file directly without converting to pdf).
   Done
   20% of total points will be taken off if run examples are not representative.
   20% of total points will be taken off if run examples are not clearly marked/annotated.

3. A discussion page
   (a) Hardware and software used by your program,
   I completed this assignment using my personal computer with PyCharm Professional Version: 2023.2.1.

   (b) Features of your program, e.g., data structures, algorithms, programming styles, etc.
   This program combines data analysis, machine learning, and data visualization techniques to analyze datasets. It allows users to select a dataset and performs either linear regression or logistic regression analysis depending on the function called. The program employs exception handling to ensure user input validity and provides informative output for analysis results.

   (c) Problems you encountered during your work, and
   None

   (d) Assigned discussion problems, if there is any.
   The assigned discussions were answered.

   (e) Fill in the following table and submit it along with your above submissions.

Total (approximate) time spent on the assignment	20 hours	Total (approximate) time for the correction part	1 hour
Problems and difficulties encountered	Can be confusing		
Reflections (good and bad) on the assignment	None		

Any comments and suggestions	None

20% of total points will be taken off if no discussion page is submitted.