Kristy Nguyen
NSHE: 5006243601
Thomas Bryant
NSHE: 2000193948

CS 422-1001
Assignment #5

Assignment 5 Summary Report

Dataset

We used the **`load_diabetes`** dataset that is built-in and provided by the scikit-learn library. The dataset contains ten baseline variables, such as age, sex, body mass index, average blood pressure, and six blood serum measurements. The target variable is a quantitative measure of disease progression one year after baseline.

Source

The **`load_diabetes`** dataset is derived from a study conducted by Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani.

Characteristics

The characteristics are as follows:
Number of Instances: 442
Number of Features: 10
Target Variable: Quantitative measure of disease progression
Type: Regression

Data Preprocessing Steps

The dataset was loaded using scikit-learn's **`load_diabetes`** function. The features (**`X`**) and target variable (**`y`**) were extracted from the loaded dataset. The dataset was split into training and testing sets using **`train_test_split`** with a test size of 20% and a random seed of 83. For the OLS model, we first appended the bias term of 1's both training and testing:

```
X_train = np.c_[np.ones(X_train.shape[0]), X_train]
X_test = np.c_[np.ones(X_test.shape[0]), X_test]
```

Then, we transpose and get the inverse inherently (via **`np.matmul`**):

```
XTX_inv = np.linalg.inv(np.matmul(X_train.T, X_train))
w_ols = np.matmul(np.matmul(XTX_inv, X_train.T), y_train)
```

To ensure enough iterations in the linear regression with gradient descent, we increased the number of iterations for the regressor **`max_iter=10000`** and ensured any other necessary underlying preprocessing such as regularizing the dataset, we use **`sgd_model.fit(X_train, y_train)`**.

Solution 'w' Parameter Vector

```
Solution 'w' Parameter Vector (OLS):
 [ 152.22872701  -19.07191794 -282.28783291  523.77804883
271.77759693
 -997.13750054  645.76026585  117.46978125  140.55942058
857.85358865
   70.14084646]
```

```
Solution 'w' Parameter Vector (SGD):
 [   1.30506855 -214.65335097  468.40544869  250.72081495  -
42.7720182
 -101.20883794 -224.40239312  125.97497661  405.976515
114.20671391]
```

Recent evaluation metrics for OLS (MSE, MAE, R^2)

Training Dataset

```
Results for Ordinary Least Squares (OLS) on Training Set:
Mean Squared Error: 2843.4389047402938
Mean Absolute Error: 42.951739896891546
R-squared: 0.5197660712428738
```

Test Dataset

```
Results for Ordinary Least Squares (OLS) on Test Set:
Mean Squared Error: 3004.536220845695
Mean Absolute Error: 45.0186804685078
R-squared: 0.49395645445808134
```

Recent evaluation metrics for linear regression with gradient descent (MSE, MAE, R^2)

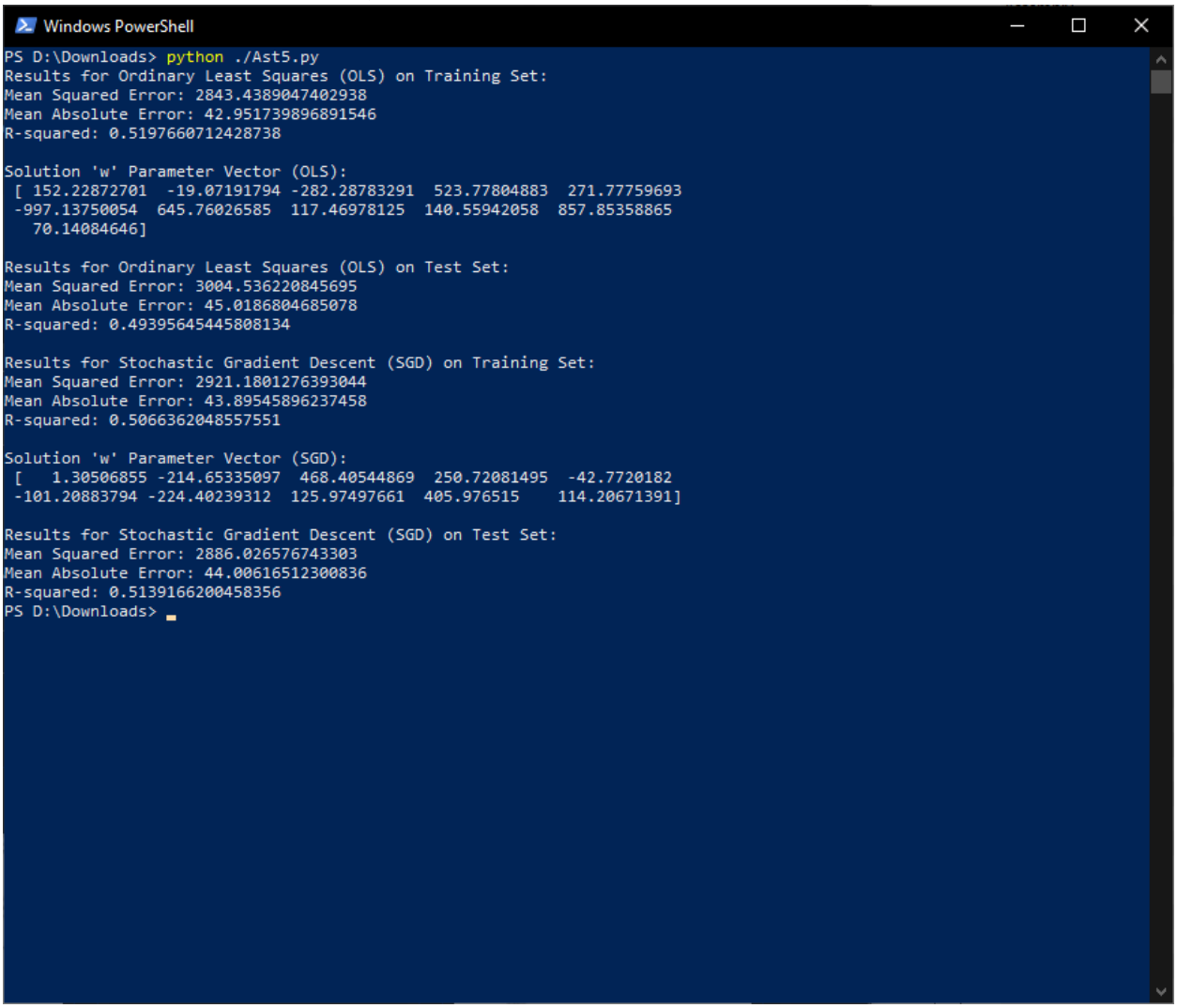
Training Dataset

```
Results for Stochastic Gradient Descent (SGD) on Training Set:
Mean Squared Error: 2921.1801276393044
Mean Absolute Error: 43.89545896237458
R-squared: 0.5066362048557551
```

Test Dataset

```
Results for Stochastic Gradient Descent (SGD) on Test Set:
Mean Squared Error: 2886.026576743303
Mean Absolute Error: 44.00616512300836
R-squared: 0.5139166200458356
```

Example Execution of Code (To compare with OLS scratch Results)



Example Execution with OLS from Scratch