Linear Regression on Diabetes Dataset

Name: Ayush Pramod Chavan
Roll No: 23
PRN: 12310917

A. What is Linear Regression? Linear Regression is a fundamental supervised learning algorithm used for predicting a continuous dependent variable (Target, y) based on one or more independent variables (Features, X). In the context of Deep Learning, Linear Regression can be viewed as a Single-Layer Neural Network with input layer (X), weights (w), bias (b), and a linear (identity) activation function.

B. Mathematical Equation The relationship is modeled as a straight line (or hyperplane in higher dimensions):

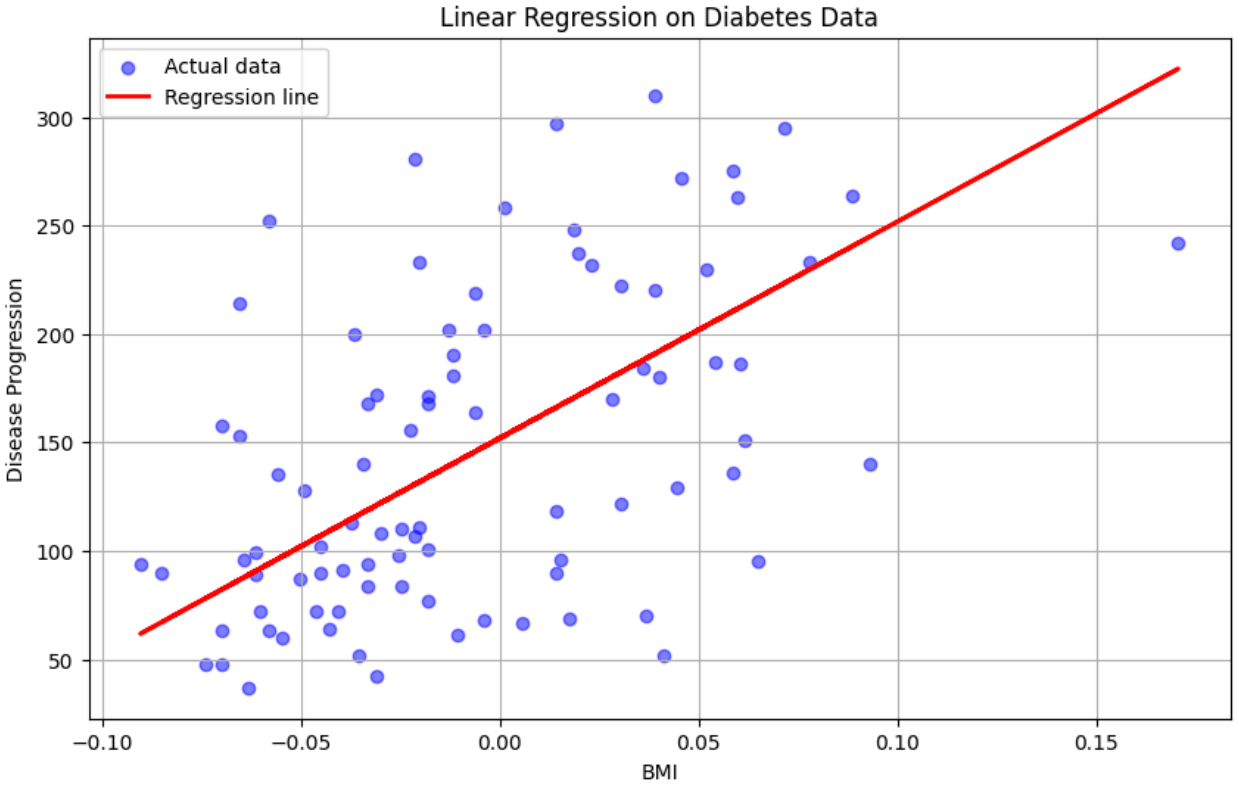
$y = wX + b$

Where y is the predicted value, X is the input feature, w is the weight (slope), and b is the bias (intercept).

In [ ]:
```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error
from sklearn.datasets import load_diabetes
import pandas as pd

# 1. Load the Diabetes dataset
diabetes = load_diabetes()
df = pd.DataFrame(diabetes.data, columns=diabetes.feature_names)
df['target'] = diabetes.target


X = df[['bmi']]
y = df['target']


X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, randc


lin_reg = LinearRegression()
lin_reg.fit(X_train, y_train)


y_pred = lin_reg.predict(X_test)
```

```
mse = mean_squared_error(y_test, y_pred)
rmse = np.sqrt(mse)

print(f"Mean Squared Error (MSE): {mse}")
print(f"Root Mean Squared Error (RMSE): {rmse}")
print(f"Model R^2 Score: {lin_reg.score(X_test, y_test):.4f}")
# 6. Visualize the results
plt.figure(figsize=(10, 6))
plt.scatter(X_test, y_test, color='blue', label='Actual data', alpha=0.5)
plt.plot(X_test, y_pred, color='red', linewidth=2, label='Regression line')
plt.title('Linear Regression on Diabetes Data')
plt.xlabel('BMI')
plt.ylabel('Disease Progression')
plt.legend()
plt.grid(True)
plt.show()
```

```
Mean Squared Error (MSE): 4061.8259284949268
Root Mean Squared Error (RMSE): 63.73245584860925
Model R^2 Score: 0.2334
```



In [ ]:
```
print(df.head())
```

```
        age       sex       bmi        bp        s1        s2        s3  \
0  0.038076  0.050680  0.061696  0.021872 -0.044223 -0.034821 -0.043401
1 -0.001882 -0.044642 -0.051474 -0.026328 -0.008449 -0.019163  0.074412
2  0.085299  0.050680  0.044451 -0.005670 -0.045599 -0.034194 -0.032356
3 -0.089063 -0.044642 -0.011595 -0.036656  0.012191  0.024991 -0.036038
4  0.005383 -0.044642 -0.036385  0.021872  0.003935  0.015596  0.008142

         s4        s5        s6  target
0 -0.002592  0.019907 -0.017646   151.0
1 -0.039493 -0.068332 -0.092204    75.0
2 -0.002592  0.002861 -0.025930   141.0
3  0.034309  0.022688 -0.009362   206.0
4 -0.002592 -0.031988 -0.046641   135.0
```

Code Explanation

1. **Load Dataset**: The code begins by loading the Diabetes dataset from `sklearn.datasets`. This dataset contains physiological data from diabetes patients.
2. **Data Preparation**: The data is converted into a pandas DataFrame for easier manipulation. We select 'bmi' as our feature (X) and 'target' (disease progression) as our target variable (y).
3. **Train-Test Split**: The dataset is split into training and testing sets. 80% of the data is used for training the model, and 20% is used for testing its performance.
4. **Model Training**: A `LinearRegression` model is instantiated and trained on the training data using the `fit` method.
5. **Prediction and Evaluation**: The trained model is used to make predictions on the test set. The performance of the model is evaluated using Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the R^2 score.
6. **Visualization**: The results are visualized by plotting the actual data points and the regression line predicted by the model.

Conclusion

The linear regression model was trained on the Diabetes dataset to predict disease progression based on BMI.

The model's performance on the test set is as follows:

- **Mean Squared Error (MSE):** 4061.83
- **Root Mean Squared Error (RMSE):** 63.73
- **R^2 Score:** 0.2334

The R^2 score of 0.2334 indicates that approximately 23.34% of the variance in the disease progression can be explained by the BMI. While the model has learned a linear relationship, the low R^2 score suggests that BMI alone is not a strong predictor of disease progression in this dataset, and a more complex model or additional features might be needed for better performance. The scatter plot visually confirms this, showing a weak positive linear relationship.