Amirreza Akbari | 810899045
Reza Baghestani | 810899046
Hananeh Jamali | 810899053

Purpose:

The purpose of this assignment is to explore various regression techniques and evaluate their performance in predicting diabetes progression using the Diabetes dataset from the scikit-learn library. We aim to implement fundamental regression functions, build a linear regression model, and assess its effectiveness in predicting disease progression.

Summary:

In this report, we will analyze the process of predicting diabetes progression using regression analysis techniques. We will first explore the Diabetes dataset, preprocess the data, and split it into training and testing sets. Then, we will implement key regression functions from scratch, build a linear regression model, and evaluate its performance using various metrics.

Dataset Used:

The Diabetes dataset consists of 442 instances with ten baseline variables, including age, sex, body mass index (BMI), average blood pressure (BP), and six blood serum measurements. The target variable is a quantitative measure of disease progression one year after baseline. We use this dataset for both training and testing purposes to predict diabetes progression.

Warm-Up:

The Warm-Up section serves as an initial preparatory step to familiarize with the dataset and perform basic data preprocessing tasks.

Steps 1, 2 and 3: Data Loading and Exploration: The dataset is loaded from the scikit-learn library using the load_diabetes() function. The head() function from pandas is employed to display the first ten rows of the dataset, providing a glimpse into its structure. Additionally, the dtypes attribute is utilized to print the data types of each feature, ensuring they are numeric.

Step 3: Handling Missing Values: The code checks for missing values in the dataset using the isnull() function, followed by sum() to compute the total number of missing values. If missing values are detected, they are replaced with NaN values. This ensures data integrity and prevents potential issues during analysis.

Step 4: Feature Normalization: Feature normalization is performed to ensure that all features are on a similar scale, preventing certain features from dominating others in the analysis. The StandardScaler() function from scikit-learn is utilized to standardize the features by removing the mean and scaling to unit variance.

Step 5: Data Splitting: The dataset is split into training and testing sets using the train_test_split() function from scikit-learn. This step ensures that the model's performance is evaluated on unseen data, providing a more accurate assessment of its generalization ability.

Step 6: Number of Instances Confirmation: Finally, the number of instances in both the training and testing datasets is displayed to confirm the successful split. This step ensures that sufficient data is available for training and testing the regression model.

Main Task:

The Main Task section focuses on building and evaluating a linear regression model to predict diabetes progression based on the dataset's features.

Part 1: Functions' Implementation: The first part involves implementing fundamental regression functions, including Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 Score (Coefficient of Determination). These functions are essential for evaluating the performance of the regression model.

Part 2: Building and Training the Linear Regression Model: The next step involves building a Linear Regression model from scratch. The LinearRegressionFromScratch class is defined, which includes methods for fitting the model to the training data (fit()) and making predictions (predict()). Alternatively, the scikit-learn LinearRegression model can be utilized for comparison.

Part 3: Model Evaluation: In this part, the performance of the regression model is evaluated using various metrics. Predicted values are compared with actual progression measures using scatter plots for both training and testing datasets. Additionally, evaluation metrics such as MSE, MAE, RMSE, and R^2 score are calculated for both training and testing data to assess the model's accuracy and generalization ability.

Part 4: Ordinary Least Squares (OLS): Finally, the Ordinary Least Squares (OLS) method is employed to fit the linear regression model using the training data. The sm.OLS function from the statsmodels library is utilized for this purpose. Model statistics, including coefficients, p-values, and summary statistics, are obtained and analyzed to gain insights into the relationship between the features and the target variable.

Questions:

Question 1) Analysis and evaluation of the results of the linear regression model

1. **Mean Squared Error (MSE)**:
   - MSE measures the average squared difference between the predicted values and the actual values. Lower MSE indicates better model performance, as it means the model's predictions are closer to the actual values.
   - In this case: The MSE values for both training and testing data are relatively low, which is a positive sign. It indicates that the model's predictions are close to the actual values on average.
2. **Mean Absolute Error (MAE)**:
   - MAE measures the average absolute difference between the predicted values and the actual values. Similar to MSE, lower MAE indicates better model performance.
   - In this case: The MAE values for both training and testing data are relatively low, indicating that the model's predictions have small errors compared to the actual values.
3. **Root Mean Squared Error (RMSE)**:
   - RMSE is the square root of the MSE and provides a measure of the spread of prediction errors. Like MSE and MAE, lower RMSE indicates better model performance.

   - In this case: The RMSE values for both training and testing data are also low, suggesting that the model's predictions have small variability from the actual values.

4. **R² Score (Coefficient of Determination)**:
   - R² score represents the proportion of the variance in the dependent variable that is predictable from the independent variables. It ranges from 0 to 1, where 1 indicates a perfect fit.
   - In this case: The R² scores for both training and testing data are moderate, indicating that the model explains a reasonable amount of the variance in the dependent variable.

Since the values of the target variable (y) has been normalized and a large portion of the values are between -1 and 1, the values of MSE, MAE and RMSE increase in the mentioned order.

Question 2) R² and Adjusted R² values

The R-squared (R²) and Adjusted R-squared (Adj. R²) values are statistical measures used to assess the goodness of fit of a regression model to the observed data. Here's what these values indicate and their implications:

1. **R-squared (R²):**
   - R-squared is a measure of how well the independent variables explain the variation in the dependent variable. It represents the proportion of the variance in the dependent variable that is explained by the independent variables in the model.
   - R-squared values range from 0 to 1, where 0 indicates that the model does not explain any of the variability in the dependent variable, and 1 indicates that the model explains all of the variability.
   - Higher R-squared values indicate a better fit of the model to the data. However, high R-squared values do not necessarily imply that the model has predictive power or that it is the best model for making predictions.
2. **Adjusted R-squared (Adj. R²):**
   - Adjusted R-squared is a modified version of R-squared that adjusts for the number of predictors in the model. It penalizes the addition of unnecessary predictors that do not improve the model's fit, preventing overfitting.
   - Like R-squared, Adj. R² values range from 0 to 1, and higher values indicate a better fit of the model to the data.
   - Adj. R² is often considered a more reliable measure of model fit, especially when comparing models with different numbers of predictors.

Implications of High or Low Values:

- High R-squared and Adj. R² values suggest that a large proportion of the variability in the dependent variable is explained by the independent variables, indicating a good fit of the model to the data.
- Low R-squared and Adj. R² values suggest that the independent variables in the model do not effectively explain the variation in the dependent variable, indicating a poor fit of the model to the data.

However, it's essential to interpret R-squared and Adj. R² values in the context of the specific problem and the nature of the data. A high R-squared does not necessarily imply a useful or accurate model, and a low R-squared does not always indicate a poor model.

Adjusted R-squared is preferred when comparing models with different numbers of predictors, as it accounts for model complexity.

In this case: The R-squared value of approximately 0.514 indicates that the model explains about 51.4% of the variability in the dependent variable (disease progression measure). This suggests a moderate degree of predictive power. The Adjusted R-squared value, approximately 0.502, considers model complexity and is slightly lower, indicating potential overfitting or non-significant predictors. Nonetheless, it still suggests a reasonable fit to the data.

Question 3) P-values

P-values are an important measure in statistical hypothesis testing, particularly in the context of linear regression analysis. They indicate the probability of observing the data or more extreme data if the null hypothesis is true. In the case of linear regression, the null hypothesis typically states that there is no relationship between the predictor variable (feature) and the response variable (target).

Interpretation for the p-values obtained for each feature:

1. **Age (age):** The p-value is 0.835903, indicating that there is no significant relationship between age and the progression of diabetes. This means that age may not be a strong predictor of diabetes progression in the model.
2. **Sex (sex):** The p-value is 5.49142e-05, which is very small. This suggests that there is a significant relationship between sex and diabetes progression.
3. **Body Mass Index (bmi):** The p-value is 7.9048e-14, indicating a highly significant relationship between BMI and diabetes progression. BMI is likely to be a strong predictor of diabetes progression in the model.
4. **Average Blood Pressure (bp):** The p-value is 4.3803e-07, suggesting a significant relationship between blood pressure and diabetes progression.
5. **Six blood serum measurements (s1, s2, s3, s4, s5, s6):** These features have varying p-values. Generally, lower p-values indicate a more significant relationship with diabetes progression. For example, s5 has a very small p-value (5.97732e-05), indicating a significant relationship, while s3 has a higher p-value (0.520235), suggesting a weaker relationship.

An appropriate threshold for p-values depends on the significance level chosen for the hypothesis test. Commonly used significance levels include 0.05 (5%) and 0.01 (1%). Features with p-values below the chosen significance level are considered statistically significant predictors in the model.

Based on the provided p-values:

- Features such as sex, bmi, bp, and s5 have p-values below 0.05, suggesting they are likely to be statistically significant predictors.
- Features with higher p-values, such as age, s2, s3, s4, and s6, may have weaker or non-significant relationships with diabetes progression in the model.

Question 4) Important Features

Same as the results of previous question (based on p-values):

Overall, based on the p-values obtained from the OLS regression analysis, age may not be a significant predictor of diabetic condition in this dataset. However, variables such as sex, BMI, and blood pressure appear to be highly significant predictors. Further analysis and interpretation of these results may provide insights into the relative importance of each feature in predicting an individual's diabetic condition.