**Regression Analysis Report: Diabetes Dataset**

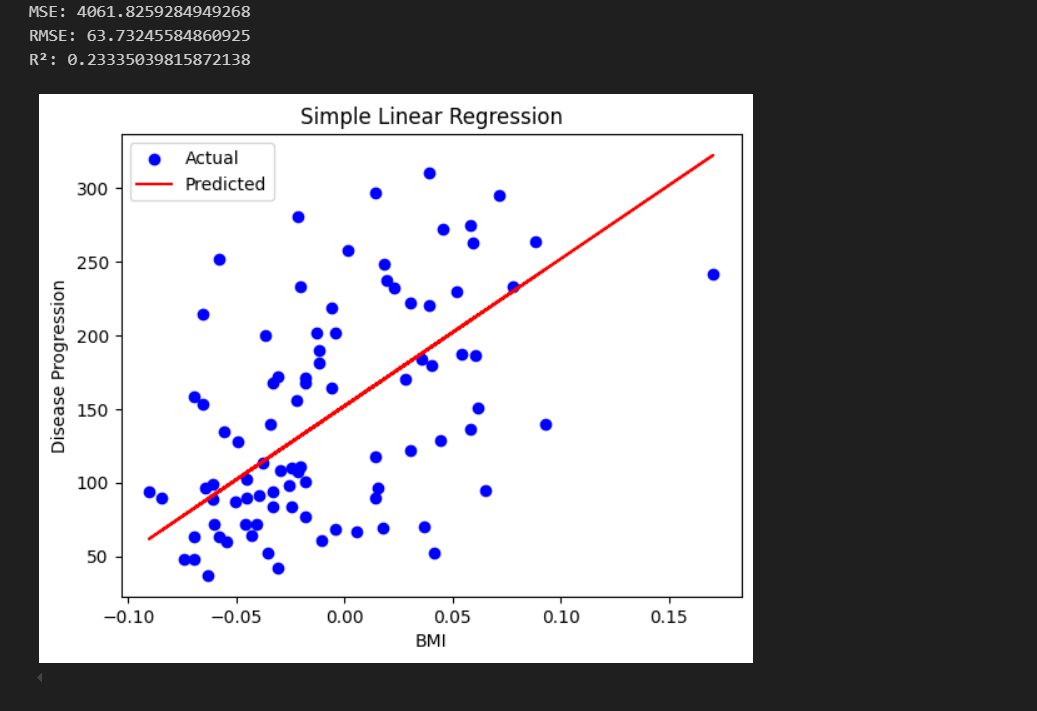
1. Introduction

This study offers a comprehensive review of regression methods that were used with the scikit-learn Diabetes dataset. The primary goal of this lab is to compare the accuracy of many regression models in predicting the course of illness. These models include Linear, Multiple, Polynomial, Ridge, and Lasso Regression. Additionally, regularization's significance in preventing overfitting is highlighted (Olisah et al., 2022). Our assessment metrics of choice include R-squared, Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Squared Error (MSE). We next evaluate and contrast these models. You can see the accuracy and data fit of each model's predictions visually.

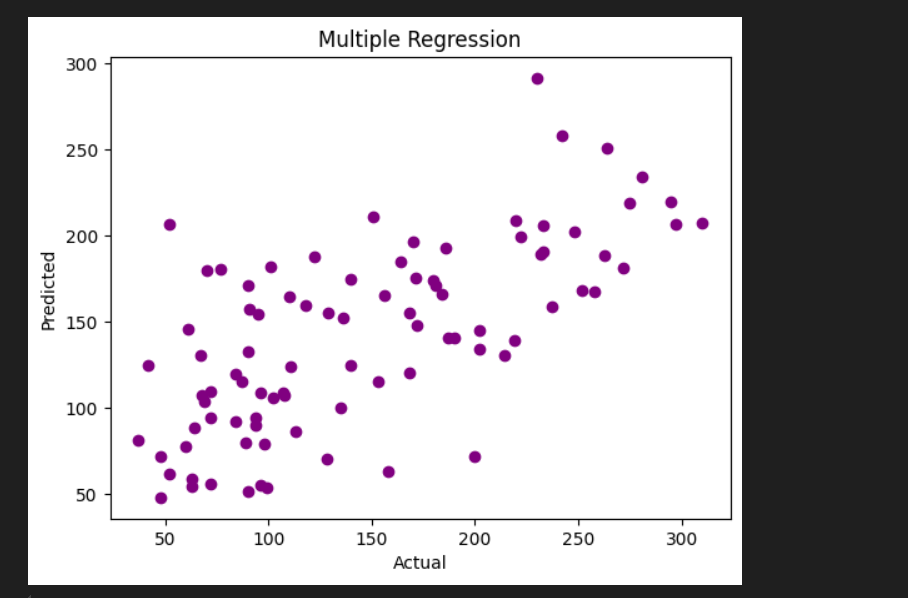
**2. Data Preparation**

There are 442 samples in the Diabetes dataset, and each one has 10 standardized characteristics that stand in for clinical factors including age, body mass index, blood pressure, and other serum values. A quantitative assessment of illness progression one year after baseline is reflected in the target variable (Olisah et al., 2022). No outliers or missing values were found during preparation since the dataset has previously been cleaned and standardized. Because of this, we could compare models on a consistent foundation without having to do any more transformations.

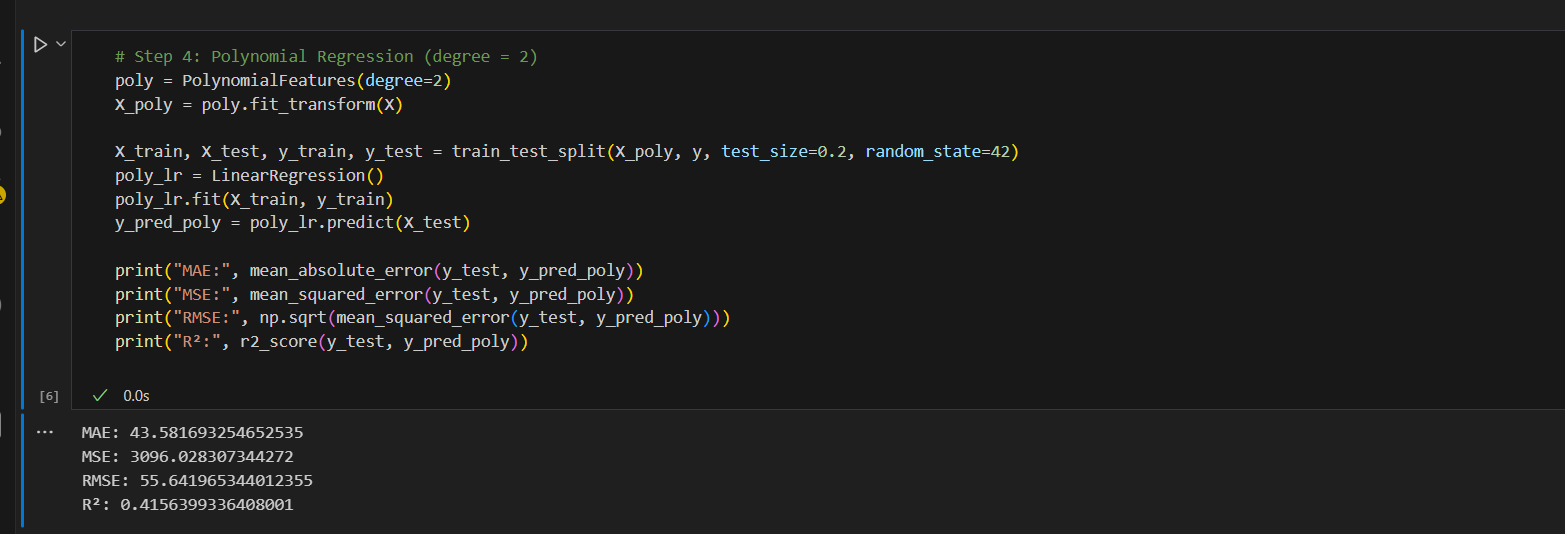
**3. Simple Linear Regression**

The 'bmi' characteristic was used as the only independent variable in a Simple Linear Regression model. Despite providing a clear picture of the connection between BMI and illness development, this method only explained around 23% of the target variable's variation (R² = 0.233). This proved that it is impossible to capture the complexities of illness development using only one attribute (Olisah et al., 2022). The low predictive potential of the model is further shown by the error metrics (RMSE = 63.73, MAE = 52.26). This method was shown to be underfitting when the projected values were plotted against the actual values.

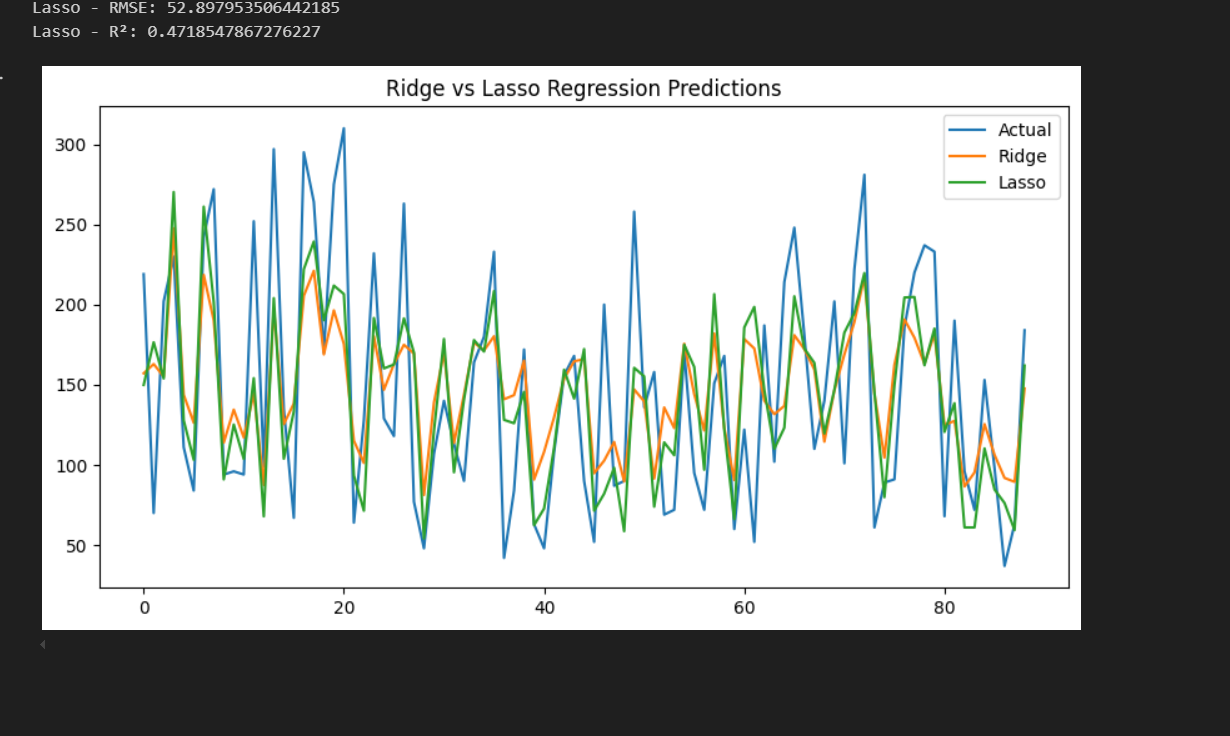
**4. Multiple Regression**

Multiple Regression considerably improved performance with a R² of 0.453 after expanding the model to include all 10 characteristics. This provides further evidence that using a combination of clinical markers improves the accuracy of disease progression forecasts (Olisah et al., 2022). The simpler model showed less generalizability, as seen by the lower error metrics (RMSE = 53.85, MAE = 42.79). The more closely the actual and projected values cluster along the diagonal of a scatter plot comparing the two, the better the model fit.

**5. Polynomial Regression**

In order to capture nonlinear interactions among variables, a 2-degree Polynomial Regression was used. Despite the model's inclusion of interaction and quadratic factors, which increased complexity, the Multiple Linear model outperformed it with a R² value of 0.416. Overfitting may also be occurring, as shown by a little rise in RMSE (55.64) (Olisah et al., 2022). Although polynomial features might improve performance on some datasets, they had no discernible effect on the Diabetes dataset, suggesting that linear correlations could be enough for representing this specific prediction job.

**6. Regularization with Ridge and Lasso Regression**

The use of Ridge and Lasso Regression methods helped avoid overfitting. By using Ridge Regression with α=1.0, overfitting was somewhat reduced, leading to a R² value of 0.419. The highest overall result was achieved using Lasso Regression with α=0.1, which had a R² value of 0.472 and the lowest RMSE of 52.90 (Olisah et al., 2022). Both interpretability and robustness are enhanced by Lasso's capacity to reduce coefficient sizes and remove superfluous features. By comparing the Lasso and Ridge predictions visually, we can see that Lasso consistently produces accurate results, which suggests that it strikes a good compromise between complexity and precision.

**7. Conclusion**

When compared to single-variable models, the predictive power of multiple regression approaches is much higher, according to the regression analysis. As an example of how regularization may improve prediction and model simplicity, Lasso Regression performed better than the others (Olisah et al., 2022). It seems that the Diabetes dataset does not gain much from nonlinear modeling, since polynomial regression adds complexity without substantially improving outcomes. Lasso Regression is the best overall solution for this situation in terms of forecasting how the sickness will proceed.

**References**

Olisah, C. C., Smith, L., & Smith, M. (2022). Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective. Computer Methods and Programs in Biomedicine, 220, 106773.