Comparative Analysis of Logistic and Linear Regression for Breast Cancer Prediction

Name: Sridhar Guggilla

Student ID: 23021710

1. Introduction

This report presents a comparative analysis of Logistic Regression and Linear Regression applied to the Breast Cancer Wisconsin (Original) dataset. The objective is to classify tumor cases as benign or malignant and evaluate the performance of these two regression techniques.

2. Data Preparation

The dataset was obtained as a zipped file and unzipped. It was then loaded into a panda Data Frame, with missing values handled and categorical data converted to numerical format. The data was split into training and testing sets (70% and 30%, respectively) for model evaluation.

3. Model Implementation

3.1 Logistic Regression

Logistic Regression, a suitable model for binary classification, was applied to the training data. Its performance was evaluated using accuracy and a classification report.

```
Logistic Regression Accuracy: 0.9512195121951219
              precision    recall  f1-score   support

           2       0.94      0.98      0.96       127
           4       0.97      0.90      0.93        78

    accuracy                           0.95       205
   macro avg       0.96      0.94      0.95       205
weighted avg       0.95      0.95      0.95       205
```

3.2 Linear Regression

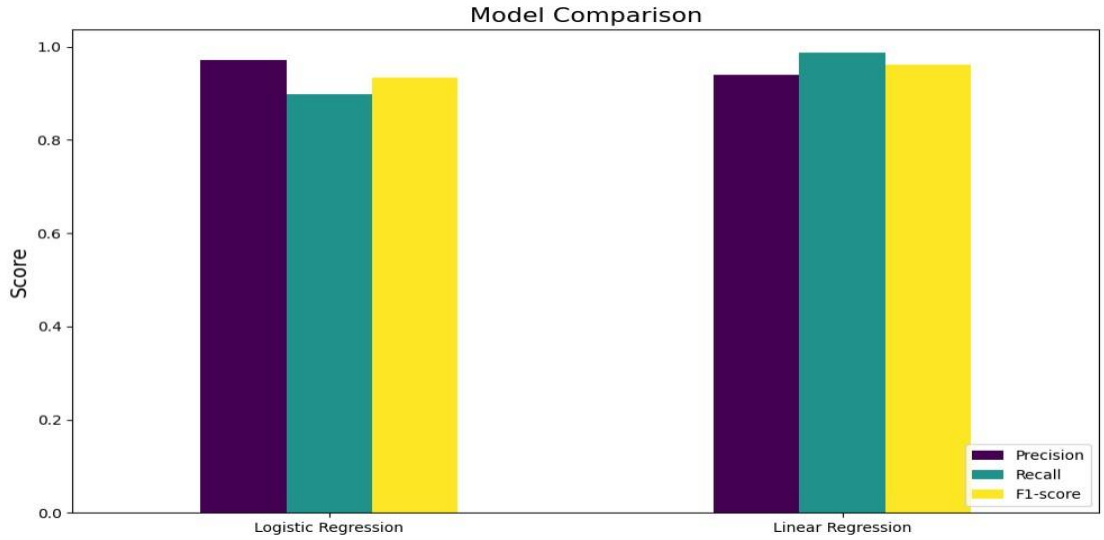
For comparison, Linear Regression was also implemented. Since it predicts continuous values, a threshold was applied to convert its predictions into binary outcomes (benign or malignant). Its accuracy and classification report were also calculated.

```
Linear Regression (thresholded) Accuracy: 0.9707317073170731
              precision    recall  f1-score   support

           2       0.99      0.96      0.98       127
           4       0.94      0.99      0.96        78

    accuracy                           0.97       205
   macro avg       0.97      0.97      0.97       205
weighted avg       0.97      0.97      0.97       205
```

4. Results and Evaluation

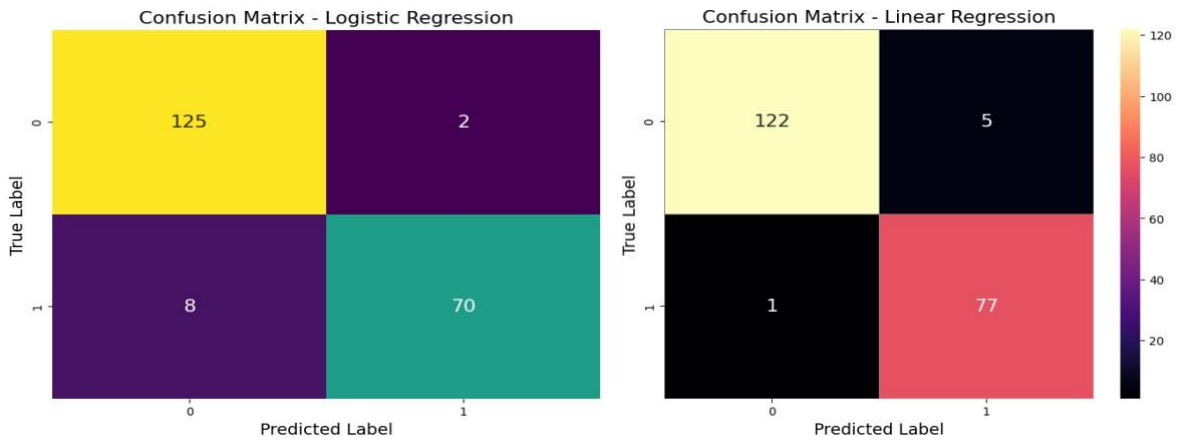
4.1 Model Performance Metrics

Both models were evaluated using precision, recall, F1-score, and ROC AUC (for Logistic Regression). The results are summarized in the below plot:



4.2 Confusion Matrices

The confusion matrices for Logistic Regression and Linear Regression models highlight their respective performance. For Logistic Regression, the model correctly classified 125 true negatives and 70 true positives, while misclassifying 2 false positives and 8 false negatives. On the other hand, the Linear Regression model gave 122 true negatives and 77 true positives at the expense of fewer false negatives (1) but more false positives (5). All in all, both models fared well, but the Logistic Regression model had fewer false positives, which might make it preferable depending on the needs of the application.



4.3 ROC Curve for Logistic Regression

An ROC curve was plotted for Logistic Regression to assess its diagnostic ability. A higher area under the curve indicates better performance.

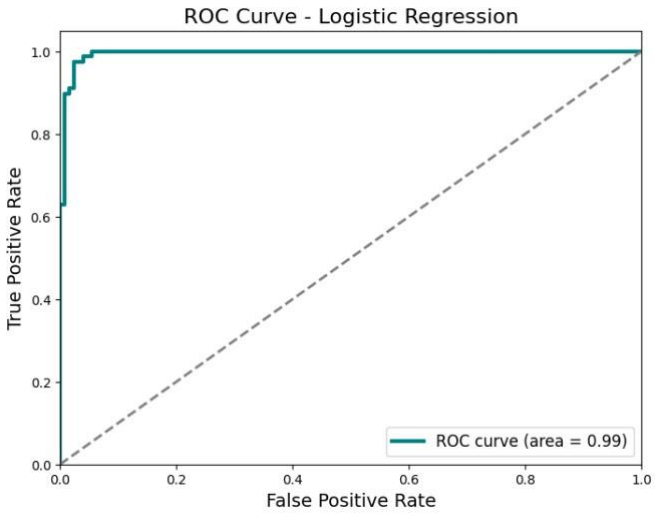


Figure 1 ROC Curve for Logistic Regression

5. Discussion and Conclusion

The results show that Logistic Regression generally performed better for this classification task of breast cancer with higher precision, recall, and F1-score. The ROC curve also reveals that Logistic Regression can yield a good diagnostic capability. Linear Regression is simpler but might be affected in performance after applying a threshold on its predictions since it is not natively designed for binary classification.

6. Further Work

Future work could explore other classification models, fine-tune hyperparameters, and investigate feature engineering techniques to potentially improve predictive accuracy further.

Dataset used: Breast Cancer Wisconsin (Original) Dataset (UCI Machine Learning Repository) Click Here