Analysis of Logistic Regression and Linear Regression on a Breast Cancer Dataset

Name: Vishakanan Sivarajah
Student ID: 23090811
GitHub Link: https://github.com/Vishakanan/Data-Mining-Assignment

Introduction

One of the primary causes of death for women is breast cancer. Accurate classification is essential for prompt diagnosis and therapy. The distinction between benign and malignant tumors is aided by machine learning techniques.
Using the Wisconsin Breast Cancer dataset, this study assesses both linear regression and logistic regression. Logistics Regression serves as a classifier, whereas Linear Regression predicts classification probabilities. The objective is to evaluate their overall performance, accuracy, and precision.

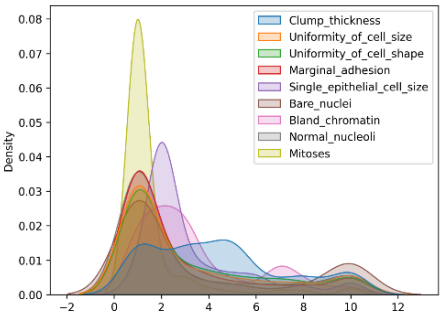
Methodology

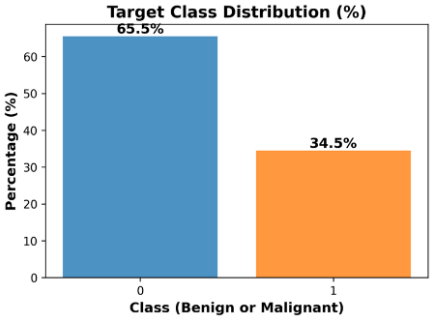
There are eleven attributes in the Wisconsin-sourced Breast Cancer dataset. Except for the ID and class attributes, all values range between 1 and 10. There are two possible values for the target variable: 2 for benign and 4 for malignant. This study applies Logistic Regression for classification and Linear Regression to model probability estimates. Logistic Regression is trained to classify breast cancer cases as benign or malignant, predicting class labels and malignancy probabilities. Model performance is evaluated using accuracy scores on training and test datasets. Linear Regression is then trained using the predicted malignancy probabilities from Logistic Regression as the target variable, aiming to model probability distributions. Performance is assessed using Mean Squared Error (MSE) and R^2 Score to measure how well Linear Regression fits the probability values.

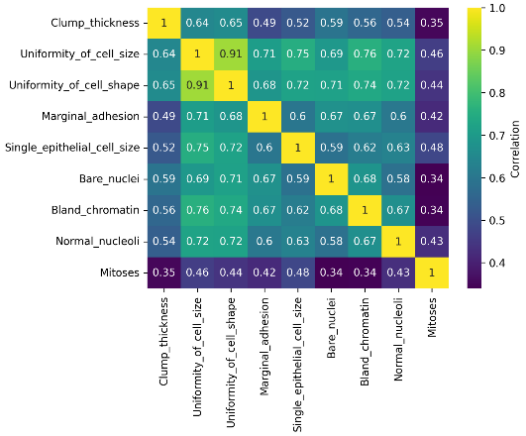
The analysis utilizes Python libraries for data preprocessing, visualization, and model training. Pandas is used for dataset handling, Scikit-learn for model implementation, and Matplotlib/Seaborn for visualization. Performance metrics such as accuracy, precision, MSE, and R^2 Score assess the models' effectiveness in classification and probability estimation.

Data Exploration

The Breast Cancer dataset is thoroughly examined during the data exploration phase, which also includes preliminary preprocessing procedures such class label mapping to binary values, mean imputation for missing values, and column renaming. A bar plot and a density plot are two examples of visualizations that provide insight on the dataset's properties.







They show that most variables are right skewed rather than normally distributed, and that most outcomes belong into the benign class. This non-Gaussian character is confirmed by the Shapiro-Wilk test. With correlation values above 0.6 for the majority of independent variables with the exception of mitoses, which exhibit weaker correlations a correlation analysis reveals a strong relationship between uniformity of cell shape and uniformity of cell size.

Data Pre-processing and Model Application

To make sure the Breast Cancer dataset is clean, organized, and appropriate for machine learning research, data pretreatment is a crucial step. After loading the dataset into a Pandas DataFrame, SimpleImputer handles missing values and, for consistency, replaces mean values. Class labels are mapped to binary values, which differentiate benign (0) and malignant (1) situations, and column names are allocated for improved readability The dataset is then split into training and testing sets using the train_test_split function from Scikit-learn, ensuring a fair evaluation of model performance.

For model application, Logistic Regression is trained on the preprocessed dataset to classify tumors as benign or malignant. The model's predictive performance is assessed using accuracy scores on both training and test sets. Additionally, Linear Regression is applied using the malignancy probabilities obtained from Logistic Regression, aiming to predict the likelihood of a tumor being malignant. The effectiveness of the Linear Regression model is evaluated using Mean Squared Error (MSE) and R^2 Score. This structured preprocessing and model training pipeline ensures a robust foundation for accurate breast cancer prediction and analysis.

Performance Evaluation

The performance of the classifiers was evaluated using classification reports for Logistic Regression and regression metrics for Linear Regression. Key evaluation metrics include accuracy, precision, recall, and F1-score for classification, and Mean Squared Error (MSE) and R^2 score for regression.

Evaluation Metrics

Model	Dataset	Accuracy	Precision (Macro Avg)	Recall (Macro Avg)	F1-Score (Macro Avg)	MSE	R^2 Score
Logistic Regression	Train Set	97%	96%	97%	97%	-	-
Logistic Regression	Test Set	97%	97%	96%	97%	-	-
Linear Regression	-	-	-	-	-	0.0152	0.9253

Comparison and Discussion

The comparison of models for breast cancer classification and probability estimation highlights key differences in their performance. Logistic Regression demonstrated strong classification capabilities, achieving an accuracy of 97% on both the training and test datasets. With a macro-average precision of 96% and recall of 97%, it effectively balances identifying true positives while minimizing false positives. The F1-score of 97% further confirms its reliable performance.

On the other hand, Linear Regression, while not a classification model, was evaluated using Mean Squared Error (MSE) and R^2 Score. The MSE of 0.0152 indicates minimal error in predicting probability values, and the R^2 score of 0.9253 suggests a strong correlation between predicted and actual malignancy probabilities. This implies that the model is effective in estimating cancer probability but may not be ideal for precise classification.

Overall, Logistic Regression emerges as the preferred model for classification due to its high accuracy and balanced performance across all metrics. However, Linear Regression can still provide valuable probability estimations that may support decision-making in medical diagnosis. The choice of model ultimately depends on whether the goal is to classify cases strictly or provide a probabilistic risk assessment.

Conclusion

In conclusion, Logistic Regression is the optimal choice for breast cancer classification due to its high accuracy and balanced performance. Linear Regression, though not suited for classification, helps estimate malignancy probabilities. These models complement each other, enhancing predictive accuracy in medical diagnostics.