Breast cancer poses a significant health challenge for women globally, underscoring the importance of early detection and diagnosis to enhance treatment outcomes. This project aims to build predictive models capable of discerning between malignant and benign breast tumours using various diagnostic features. Drawing on insights from recent advancements in machine learning applied to breast cancer detection and prognosis, the Breast Cancer Wisconsin (Diagnostic) Dataset sourced from the UCI Machine Learning Repository was used. This dataset encompasses features derived from digitised images of breast mass aspirates, including clump thickness, uniformity of cell size and shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, and mitosis. The target variable, representing tumour diagnosis, uses '2' for benign Benign 458 (65.5%) and '4' Malignant 241 (34.5%) for malignant tumours.

The dataset comprises 699 samples, with each sample containing information on 10 features and 1 class label. Notably, the dataset includes 16 instances with missing feature values denoted as '?'[2]. In selecting linear and logistic regression models, simplicity, interpretability, and effectiveness were prioritised for classification tasks. Linear regression offered insights into feature-target relationships, while logistic regression facilitated binary classification, predicting tumour malignancy probability. Both models provided a balance between performance and interpretability, ideal for this breast cancer classification project.

The Breast Cancer Wisconsin dataset was accessed through the **ucimlrepo** Python package and then data preprocessing steps were conducted to handle missing values and standardise feature scales. Exploratory data analysis provided insights into the dataset's characteristics before constructing and evaluating both linear regression and logistic regression models for breast cancer prediction. Feature scaling ensured uniformity for numerical features, and logistic regression predicted tumour malignancy probabilities, with potential hyperparameter tuning to improve performance.

Concurrently, linear regression focused on predicting a chosen target variable, such as mean nuclear size, using the features. Evaluation metrics such as accuracy, precision, recall, F1-score for logistic regression, and MSE, R-squared for linear regression were used to assess model performance. In refining the models, hyperparameter tuning and cross-validation were employed to optimise performance, highlighting bare_nuclei as a crucial feature in both models for discerning between benign and malignant tumours.

RESULTS

Linear Regression:

Mean Absolute Error: 0.2500

Mean Squared Error: 0.1530

R-squared Score: 0.8240

Logistic Regression:

Accuracy: 0.9619

Precision: 0.9683

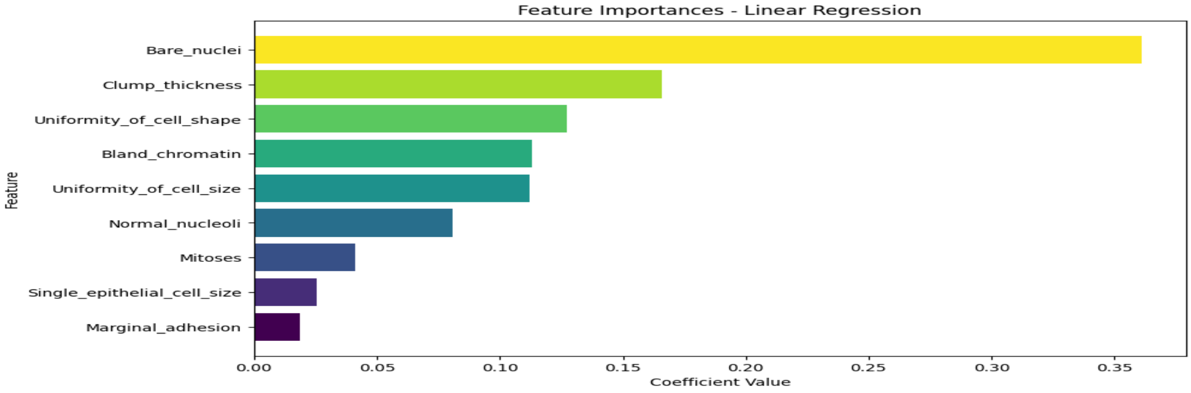
Recall: 0.9104

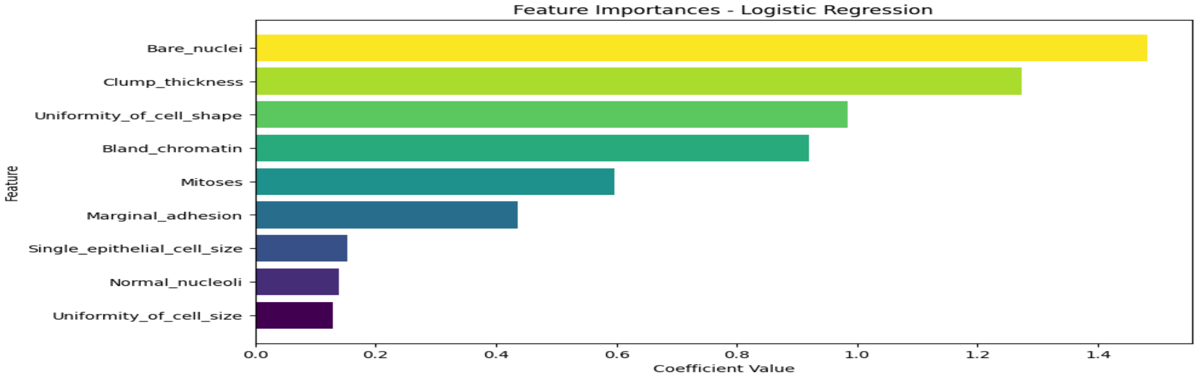
F1 Score: 0.9385

Linear Regression Cross Validation R-squared Scores: Mean of 0.8279

Logistic Regression Cross-Validation Accuracy Scores: Mean of 0.9632

The results reveal notable distinctions between the performance of linear regression and logistic regression models. For linear regression, the mean absolute error is approximately 0.25, with a mean squared error of 0.153 and an R-squared score of 0.824. Logistic regression demonstrates a high accuracy of 0.9619 (96%), along with precision, recall, and F1 scores of 0.9683, 0.9104, and 0.9385, respectively. Additionally, cross-validation further evaluates the models' performance, with linear regression achieving a mean R-squared score of 0.8279 and logistic regression attaining a mean accuracy of 0.9632. These results showcase the efficacy of logistic regression in accurately classifying breast cancer tumours, while linear regression provides insights into the continuous relationship between features and target variables. The performance of the developed models were meticulously evaluated, shedding light on their efficacy in breast cancer prediction. Logistic regression exhibited commendable accuracy, precision, recall, and F1-score metrics, showing its capability in differentiating between benign and malignant tumours with high confidence. Linear regression showcased proficiency in predicting continuous target variable values, with mean squared error (MSE) and R-squared serving as key performance indicators. Notably, bare_nuclei emerged as a critical feature in both models, underscoring its significance in driving predictive accuracy and contributing to the models' overall efficacy.





CONCLUSION

In conclusion, The objective to develop predictive models for breast cancer classification yielded promising results, with logistic regression and linear regression showcasing commendable performance. Through meticulous data preprocessing, feature engineering, and model refinement, models were successfully constructed capable of differentiating between benign and malignant tumours with high accuracy and precision. Hyperparameter tuning and cross-validation played major roles in optimising model performance, ensuring robustness and generalisation across diverse datasets. Moving forward, continued exploration and refinement of predictive modelling techniques hold promise for further enhancing diagnostic accuracy and facilitating timely interventions in the fight against breast cancer.