COMP 551 - Assignment 2 Report

Elizabeth Kourbatski

Galia Oliel-Sabbag

Colin Song

March 1, 2025

Abstract

In this assignment, we investigated the performance of linear and logistic regression machine learning models for classification. For binary classification, we used the Wisconsin Breast Cancer diagnostic data set to distinguish malignant from benign tumors. In parallel, we explore multiclass classification on the Palmer Penguins dataset, predicting penguin species based on four continuous features after appropriate pre-processing. The results indicate that logistic regression models provides a better framework for classification tasks than linear regression.

1 Introduction

In this study, we investigate the effectiveness of various regression and classification models on two distinct datasets: the Palmer Penguins dataset[1], which categorizes penguins based on their phenotypes, and the Breast Cancer Wisconsin dataset[2], which classifies tumors as malignant or benign based on their measurements.

For the Penguins dataset, a previous study employed a neural network with a softmax activation function to classify species based on culmen length, culmen depth, flipper length, body mass, and sex[3]. We evaluated both logistic regression and linear regression, achieving a perfect classification AUROC of 1.0 and accuracy of 100%, which is comparable to the neural network accuracy done in a previous study by Goldsztein and Hua's which achieved 100% accuracy[3].

For the Breast Cancer dataset, we also assessed both linear and logistic regression models. Although linear regression performed well on the Penguins dataset, it was suboptimal for cancer classification due to its assumption of a continuous relationship between features and output labels. In contrast, logistic regression outperformed linear regression, achieving an AUROC of 0.9705. Notably, our logistic regression model even surpassed previous logistic models that attained an accuracy of 96.49% using random forests[4], by focusing on a subset of only five highly predictive features rather than the full dataset.

2 Methods

2.1 Linear Regression

Linear regression is a machine learning method that models the relationship between a set of input variables/features and a continuous target variable by minimizing the sum of squared errors between predicted and actual values. The model assumes a linear combination of the features plus a bias term. Although primarily designed for regression tasks, the output can be interpreted or transformed for classification by applying a link function (e.g., sigmoid or softmax) to map the continuous predictions to probabilities.

2.2 Logistic Regression

Logistic regression is a classification algorithm that applies the logistic (sigmoid) function to a linear combination of input features, producing an output between 0 and 1 interpreted as a probability. Model parameters are learned by minimizing the cross-entropy loss, commonly using gradient-based optimization methods such as gradient descent.

2.3 Mulitclass Logistic Regression

Multiclass logistic regression generalizes logistic regression to multiple classes by replacing the sigmoid with the softmax function, which produces a probability distribution over all classes.

The model is trained to minimize the cross-entropy loss across all classes simultaneously, ensuring that the probabilities for each input sum to one. As in the binary case, gradient-based methods are typically used to optimize the model parameters.

2.4 Mulitvariate Logistic Regression

Multivariate logistic regression extends binary logistic regression to predict multiple correlated binary outcomes simultaneously. Each outcome has its own logistic function, but the outcomes share some parameters to capture potential correlations. As with other logistic models, training involves minimizing the cross-entropy loss through gradient-based methods.

3 Datasets

For the Penguin dataset, we began by dropping the "island" column and later removed the "sex" feature after chi-squared testing showed it was not a significant predictor (p=0.7657). The remaining features were each standardized. For Adelie penguins, culmen depth had a correlation of 0.2676, body mass -0.2763, flipper length -0.3447, and culmen length -0.4155. In Chinstrap penguins, culmen length correlated at 0.1791, culmen depth at 0.1280, flipper length at -0.0721, and body mass at -0.1164. Gentoo penguins showed a different pattern, with flipper length (0.4168) and body mass (0.3927) more positively correlated with the species, while culmen depth (-0.3956) was negatively correlated. These variations suggest that some features are stronger predictors for certain species than others.

The Breast Cancer dataset was used to classify tumors as either malignant or benign, with malignant tumors encoded as 1 and benign tumors as 0. Initially, no columns were dropped, and all features were standardized to ensure consistency in scale across variables. To assess feature importance, models were trained using all available features and then compared against models trained on a subset of the five most influential features, identified via simple linear regression coefficients. The top five features were *Concave_Points_3* (0.3837), *Perimeter_3* (0.3785), *Concave_Points* (0.3755), *Radius_3* (0.3754), and *Perimeter* (0.3591), which demonstrated the highest predictive power in distinguishing malignant from benign tumors.

Dataset	Number of Instances	Features	Target Parameter
Breast Cancer	569	30	Diagnosis
Penguin Dataset	342	6	species

Table 1: Summary of datasets used for this assignment.

4 Results

4.1 Breast Cancer Dataset

4.1.1 Feature Importance

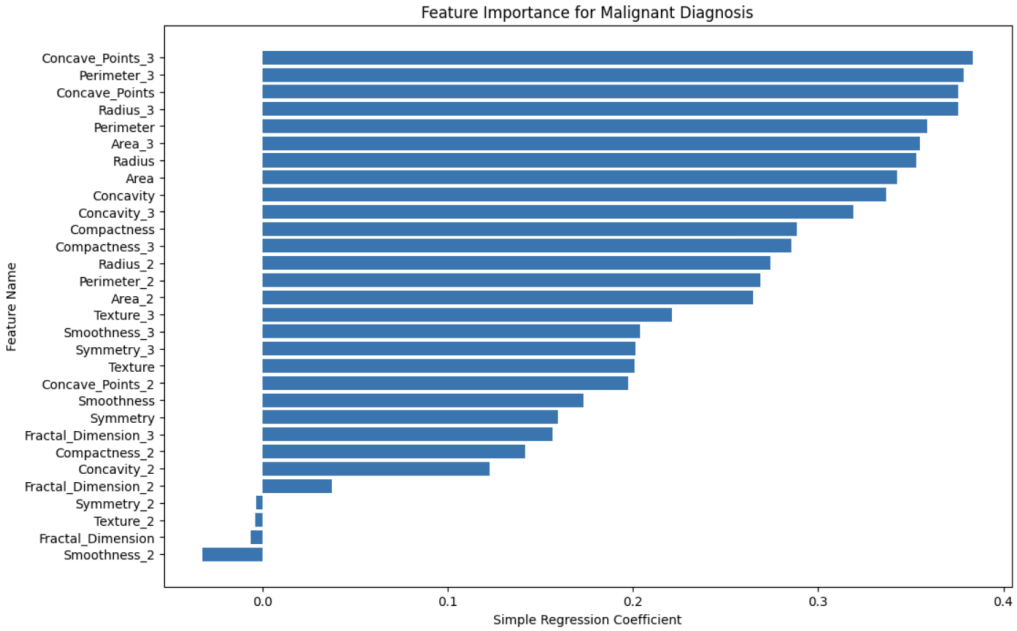

Figure 1: Feature Importance for the Breast Cancer Dataset

4.1.2 Breast Cancer Dataset Results

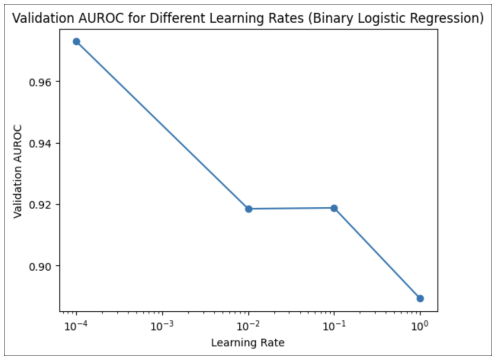

Figure 2: Validation AUROC for different learning rates using all features in the Breast Cancer dataset.

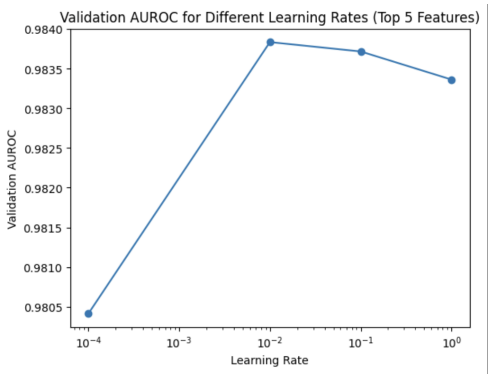

Figure 3: Validation AUROC for different learning rates using only the top 5 features in the Breast Cancer dataset.

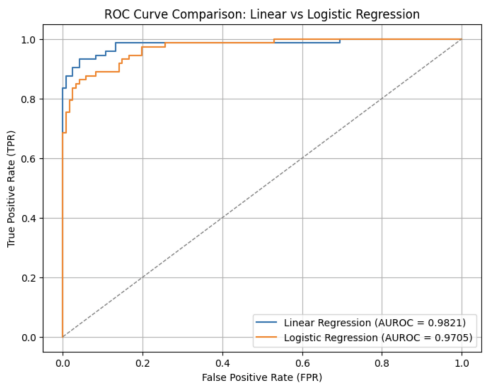

Figure 4: ROC curve comparison of Linear vs. Logistic Regression using all features in the Breast Cancer dataset.

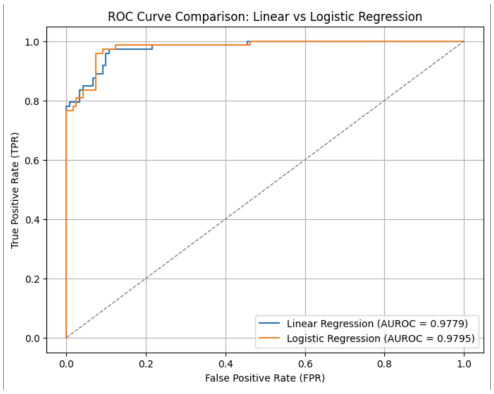

Figure 5: ROC curve comparison of Linear vs. Logistic Regression using only the top 5 features in the Breast Cancer dataset.

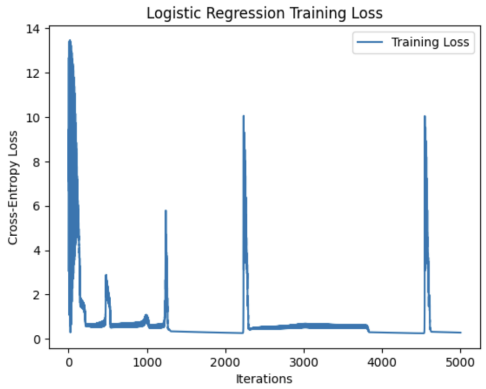

Figure 6: Logistic Regression cross-entropy training loss over iterations using all features in the Breast Cancer dataset.

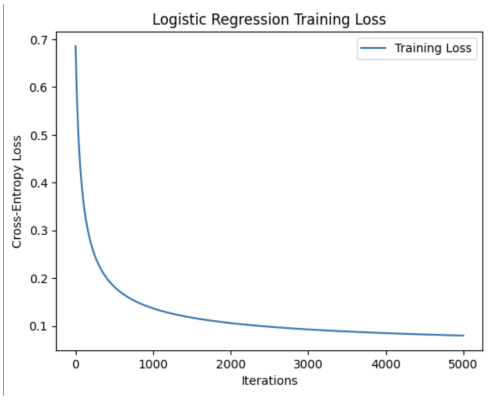

Figure 7: Logistic Regression cross-entropy training loss over iterations using only the top 5 features in the Breast Cancer dataset.

Using the full feature set of the Breast Cancer dataset, we initially identified an optimal learning rate of 10^{-4} based on validation metrics, with the final binary logistic regression model achieving a test AUROC of 0.9705 and the linear regression model obtaining an AUROC of 0.9821. However, the cross-entropy training loss was notably jittery—indicative of unstable convergence in the high-dimensional space—even after performing gradient checking with a perturbation of 1×10^{-5}. To address this, we empirically conducted feature selection and isolated the top five most influential features—*Concave_Points_3*, *Perimeter_3*, *Concave_Points*, *Radius_3*, and *Perimeter*. Retraining the models with only these features resulted in a smooth cross-entropy loss curve and shifted the optimal learning rate to 10^{-2}, with the final binary logistic regression test AUROC improving to 0.9795 and the linear regression test AUROC at 0.9779. Gradient checking in this reduced-feature scenario confirmed the accuracy of our computations, yielding a maximum relative gradient difference of 6.466928×10^{-11}.

Moreover, a comparative analysis of regression coefficients revealed that logistic regression assigns significantly higher weights to critical features (e.g., *Concave_Points_3* at 2.169, *Perimeter_3* at 1.220, and *Radius_3* at 1.109), whereas linear regression distributes coefficients more evenly, even yielding negative values for certain features such as *Area_3* (-0.249) and *Perimeter* (-0.129). This distinction underscores the superior capability of logistic regression to delineate non-linear decision boundaries essential for binary classification—a capacity that linear regression lacks.

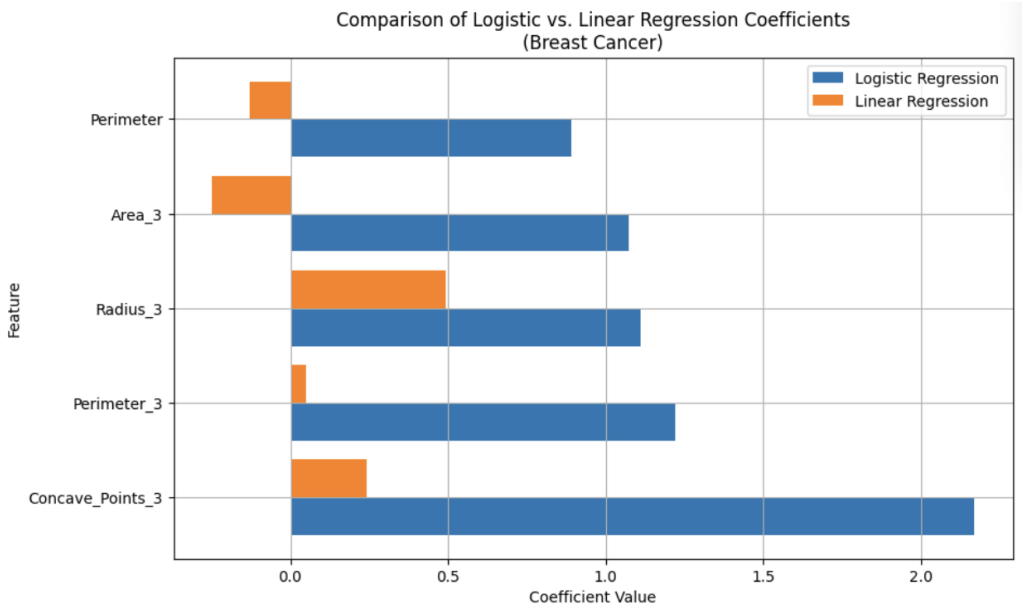

Figure 8: Comparison of logistic vs. linear regression coefficients for the Breast Cancer dataset. Logistic regression concentrates higher weights on critical features (e.g., *Concave_Points_3*), while linear regression tends to distribute coefficients more evenly, including negative coefficients for some features.

4.1.3 Extra Experiments

For the Breast Cancer dataset, our Ridge and Lasso experiments indicate that *Radius_3* and *Concave_Points_3* are the most influential features across all models. Ridge regression, which uses L2 regularization, shrinks all coefficient values without eliminating features. Lasso regression, which uses L1 regularization, also shrinks coefficients while performing feature selection by reducing certain coefficients (e.g., *Perimeter_3*) to nearly zero. By contrast, the unregularized linear regression model assigns larger coefficients, thereby increasing the risk of overfitting—particularly in datasets with correlated features. These findings suggest that Lasso is well-suited for scenarios requiring feature selection, whereas Ridge is more appropriate when retaining all features is important but controlling overfitting remains a priority.

We also fit a decision tree and a KNN model. Based on our validation metrics, the decision tree (depth = 4) achieved an AUROC of 0.8903, while KNN (5 neighbors) achieved an AUROC of 0.9657. Both performed worse than our linear and logistic models. The decision tree may have underperformed due to overfitting, whereas KNN was likely too sensitive to noisy data. Because the dataset appears to favor linear decision boundaries, linear and logistic models generalized better, leading to higher AUROC scores overall.

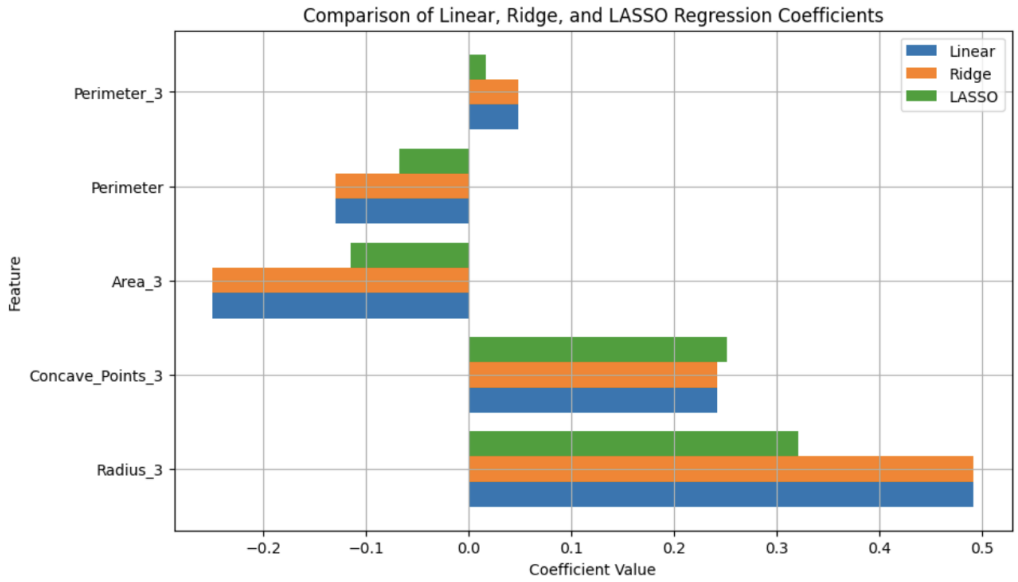

Figure 9: Comparison of linear, ridge, and Comparison of logistic vs. linear regression coefficients for the Breast Cancer dataset. Logistic regression concentrates higher weights on critical features (e.g., *Concave_Points_3*), while linear regression tends to distribute coefficients more evenly, including negative coefficients for some features.Lasso regression coefficients for the Breast Cancer dataset. Ridge (L2) shrinks coefficients without eliminating any features, whereas Lasso (L1) reduces certain coefficients (e.g., *Perimeter_3*) close to zero. Linear regression assigns larger coefficients overall, increasing the potential for overfitting.

4.2 Penguin Dataset

4.2.1 Feature Importance

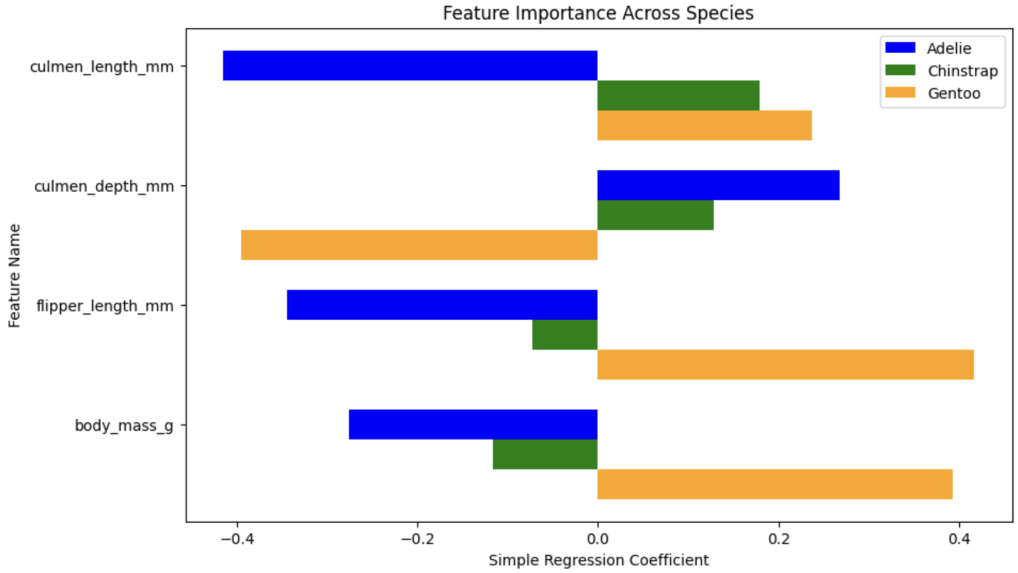

Figure 10: Feature Importance for the Penguin Dataset, all three species.

4.2.2 Penguin Dataset Results

In our experiments with the Penguin dataset, we tuned the model parameters by tracking the maximum number of iterations, tolerance (epsilon), and learning rate. After testing, the optimal configuration was determined to be an epsilon of 10^{-5}, a learning rate of 0.001, and a maximum

of 10,000 iterations. With these parameters, multiclass logistic regression achieved a perfect AUROC of 1.000 and multivariate linear regression had an accuracy of 100% on the test set. Additionally, gradient checking confirmed the correctness of our optimization process, yielding a total gradient difference of 0.00000, and the cross-entropy loss curve showed a smooth decrease.

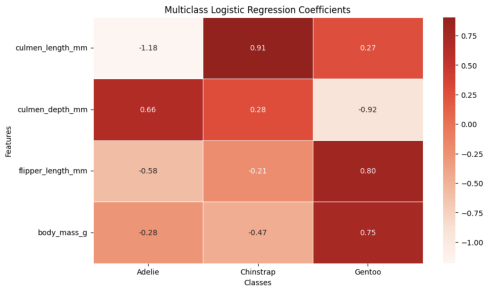

Figure 11: Heatmap visualization of multiclass logistic regression on the Penguin dataset, highlighting species-specific relationships for each feature.

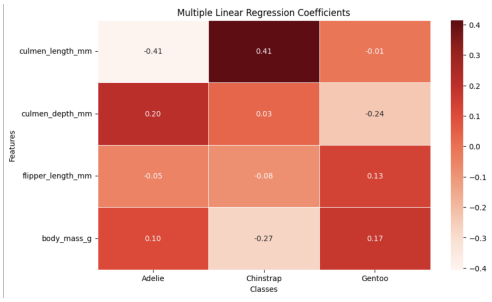

Figure 12: Heatmap visualization of multiple linear regression coefficients for the Penguin dataset, illustrating a more uniform distribution of feature weights across the three species compared to logistic regression.

A detailed comparison of the model coefficients was conducted via heatmap visualization. The multiclass logistic regression model assigns both positive and negative coefficients, thereby capturing the nuanced relationships between features and penguin species. For example, the logistic regression results indicate that *culmen length* receives a strong negative coefficient for Adelie penguins, suggesting that shorter culmen lengths are characteristic of this species, while a positive coefficient of 0.90 for Chinstrap penguins highlights longer culmen lengths as a distinguishing feature. Similarly, *culmen depth* is positively associated with Adelie (0.66) and Chinstrap (0.28) species but exhibits a strongly negative coefficient for Gentoo (-0.92). Longer flipper lengths are positively correlated with Gentoo (0.80) and negatively with Adelie (-0.59), and Gentoo penguins show the highest positive coefficient for *body mass* (0.74), indicating that heavier individuals are most likely Gentoo.

In contrast, the coefficients derived from the multiple linear regression model are more uniformly distributed across the features. For instance, the *culmen length* coefficients are relatively similar across species (-0.41, 0.41, -0.01), indicating a weaker distinction between classes. Minor variations in the coefficients for *culmen depth*, *flipper length*, and *body mass* further suggest that linear regression is less effective at separating the classes compared to logistic regression.

4.2.3 Extra Experiments

For the Penguin dataset, the Ridge and LASSO regression models yield coefficient values similar to those of linear regression, indicating that regularization does not significantly alter the relative importance of the features. This suggests that all selected features contribute meaningfully to the model. While Ridge applies L2 regularization to prevent overfitting by slightly reducing coefficient magnitudes, LASSO—using L1 regularization—slightly shrinks coefficients without eliminating any. This behavior highlights that the features in the dataset are all important predictors.

We also fitted a decision tree and a KNN model for the Penguin dataset. Using our validation metrics, the KNN model (with 3 neighbors) achieved a perfect AUROC of 1.000, while the decision tree (with a maximum depth of 2) achieved an AUROC of 0.9875. Notably, the decision tree was the only model that did not attain a perfect score.

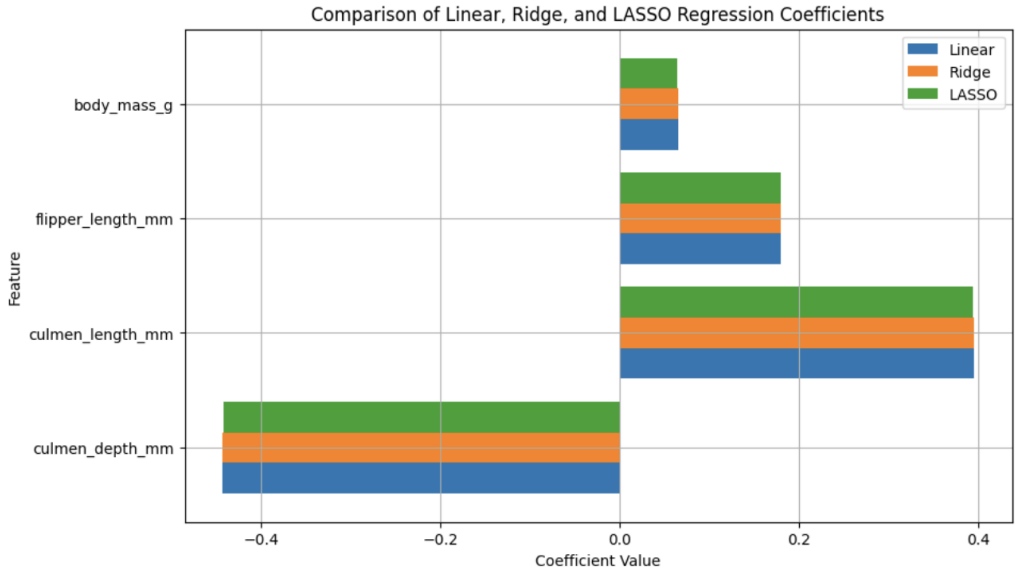

Figure 13: Comparison of linear, ridge, and Lasso regression coefficients for the Breast Cancer dataset. Ridge (L2) shrinks coefficients without eliminating any features, whereas Lasso (L1) reduces certain coefficients (e.g., *Perimeter_3*) close to zero. Linear regression assigns larger coefficients overall, increasing the potential for overfitting.

5 Discussion and Conclusion

This study investigated the performance of various regression and classification models on the Breast Cancer and Penguin datasets. Our results clearly indicate that logistic regression outperforms linear regression for classification tasks. This is primarily because logistic regression leverages the sigmoid (or softmax) function to effectively handle binary and multiclass outputs, whereas linear regression inherently assumes a continuous relationship between variables.

For the Breast Cancer dataset, logistic regression achieved a higher AUROC than linear regression. This better performance can be attributed to the use of cross-entropy loss, which is more appropriate for classification than the mean squared error (MSE) used in linear regression. Furthermore, logistic regression assigns higher absolute coefficients to critical features, hence emphasizing those that contribute most significantly to the decision boundary. In contrast, linear regression distributes weights more evenly based on feature correlations with the target variable. Features such as *Concave_Points_3*, *Perimeter_3*, and *Radius_3* emerged as key predictors, consistent with medical research that associates tumor shape and size with malignancy[5].

Similarly, in the Penguin dataset, species classification was strongly influenced by features such as culmen length, flipper length, and body mass. These findings further support the conclusion that while linear regression provides valuable insights into feature relationships, logistic regression is more effective in mapping feature importance to decision boundaries, which ensures better classification performance.

In summary, logistic regression not only offers improved predictive accuracy but also better captures the non-linear decision boundaries essential for effective classification.

6 Statement of Contribution

All team members contributed evenly to the coding and write-up portions of this project!

References

1. Penguin Dataset: https://www.kaggle.com/code/parulpandey/penguin-dataset-the-new-iris/input?select=

2. Breast Cancer Dataset: https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic

3. Using Machine Learning to Predict Penguin Species: https://www.researchgate.net/publication/371736480_Using_Machine_Learning_to_Predict_Penguin_Species

4. Breast Cancer Detection and Prevention Using Machine Learning: https://pmc.ncbi.nlm.nih.gov/articles/PMC10572157/

5. Breast Cancer Tumor Size and Shape: https://cancer.ca/en/cancer-information/cancer-types/breast/staging

6. Professor Yue Li's GitHub: https://www.cs.mcgill.ca/~yueli/teaching/COMP551_Winter2025/comp551_winter2025.html