COMP 551 Assignment 2

Jack Parry-Wingfield, Lucas Andrade, Alina Shimizu-Jozi

February 2025

Abstract

This study compares the performance of logistic regression and linear regression for classification tasks, highlighting their differences in predictive capability. While linear regression is designed for continuous target variables, logistic regression is optimized for binary classification, with multi-class logistic regression extending its applicability to categorical outcomes. Additionally, a multivariate linear regression model was applied to multi-class classification by minimizing the sum of squared errors (SSE) between predicted continuous outputs and one-hot encoded labels.

Experiments were conducted on the Breast Cancer Diagnostic (BCD) dataset for binary classification and the Palmer Archipelago Penguins (PAP) dataset for multi-class classification. Performance metrics included accuracy and area under the receiver operating characteristic curve (AUROC) and accuracy scores. Results demonstrated that logistic regression outperformed linear regression in classification tasks, achieving an AUROC of 0.9995 compared to 0.9982 for linear regression. Feature importance analysis identified "concave_points3" as the most influential feature in the BCD dataset, while "culmen_depth_mm" and "flipper_length_mm" were most significant for distinguishing penguin species. The multi-class logistic regression model achieved an accuracy of 95.80%, while multivariate linear regression attained 100% accuracy, suggesting potential over-fitting. These findings illustrate the advantages and limitations of each regression technique, emphasizing the importance of selecting appropriate models based on dataset characteristics. This comparison provides insights into the practical applications of regression models in predictive analytics.

1 Introduction

Logistic and linear regression are the most commonly used algorithms in machine learning (ML) and deep learning (DL) applications (McBride-Ellis, 2022). While linear regression is renowned for it's simplicity, interpretability and computational efficiency on linearly-separable data, logistic regression excels in binary classification tasks (Su et al., 2012; Castro & Ferreira, 2022). Further, choosing between the models largely relies on the desired output, with linear regression producing a continuous outcome variable and logistic regression giving a categorical outcome variable (Castro & Ferreira, 2022). As discussed in class, multiple logistic regression extends logistic regression to handle higher-dimensional classification tasks, while multi-class logistic regression further expands this approach to accommodate multiple possible outcomes. With applications in diabetes screening, economic impacts of trade policies and across many other fields, this model is well known for it's classification abilities (Tabaei & Herman, 2002; (Ibiyeye & Ibiyeye, 2024). Multivariate linear regression for multi-class classification is prone to over-fitting, causing a deceptively high accuracy (Hawkins, 2004). This pattern was also detectable in our implementation of the multivariate regression model. A preliminary review of the literature and lecture materials on all three models gave us insight into each algorithm's behavior before implementation. This provided accurate performance expectations during training and testing, and the ability to detect erroneous behaviour. However, to fully grasp the robustness of each model, we henceforth embark on this assignment where we write these regression models from scratch, and compare their performance on two benchmark datasets, the Palmer Archipelago penguin dataset (PAP), and the Breast Cancer Diagnostic dataset (BCD).

In Section 3, we present a series of experiments to validate the correct implementation of each model and assess their performance using standard evaluation metrics. Notably, logistic regression achieved a near-perfect AUROC of 0.9995, surpassing linear regression, which obtained a score of 0.9982. This outcome aligns with established theoretical expectations, reinforcing the suitability of logistic regression for classification tasks. Feature importance analysis revealed that in binary classification, the key features were "concave_points3," "perimeter3," and "concave_points_1," whereas for multi-class classification, "culmen_depth_mm" was most significant for Adelie and Chinstrap species, while "flipper_length_mm" was the most important feature for Gentoo. The logistic regression models for both binary and multi-class classification tasks exhibited smooth convergence, with cross-entropy loss stabilizing effectively over increasing iterations of training. In terms of accuracy, our multi-class logistic regression model attained 95.80%, while multivariate linear regression for multi-class classification perfectly fit the test data with 100% accuracy, raising concerns of possible over-fitting.

By conducting these experiments, we ensure a comprehensive analysis of each method's strengths and weaknesses, allowing us to draw conclusions on their practical applications. These findings provide a deeper understanding of when and why specific regression techniques should be used in real-world scenarios.

2 Datasets

The PAP and BCD datasets used in training and evaluating our models were publicly available through Kaggle and the UC Irvine Machine Learning repository, respectively. The BCD used for binary classification consisted of 30 features computed from digitized images of breast masses (Wolberg et al., 1993). The multivariate collection of the PAP dataset is derived from the observation of three penguin species: Chinstrap, Adélie and Gentoo in Antarctica (Pandey, 2020). From the seven columns of data, only five features were used to train the models as per instructions to remove the 'island' feature, with the last data column for species name used as the target variable. The process of data cleaning included eliminating any rows containing NaN values and duplicates, in addition to ensuring all data was standardized. Feature importance was calculated using simple regression coefficients and visualized in bar plots illustrated by figures 2, 3, 4, 5. A bar plot for each species in the PAP dataset was also visualized to see if any features were insignificant across all three classes. All features had significant importance to at least one of the species. For the BCD dataset, features below a simple regression coefficient value of 0.2 were removed to focus our training on features with higher regression coefficients, corresponding to possible increased importance (Rendall et al., 2019). Each dataset was converted into a NumPy array before training and testing the model.

3 Results

3.1 Feature Importance in Simple Linear Regression

The horizontal bar plots shown in Figures 3 and 2 represent feature importance for both binary and multi-class classification tasks respectively by plotting the regression coefficients for each feature. These coefficients indicate the strength and direction of each feature's contribution to the classification decision. A higher absolute value of a coefficient suggests a more influential feature, with positive values pushing predictions towards one class and negative values towards another (Rendall et al., 2019). The regression coefficients were obtained via simple linear regression, as specified in the assignment, allowing for an interpretable assessment of feature impact. This visualization conveys which features are most relevant in distinguishing between classes and can guide feature selection or further model refinement. A quick look at the plots show us that for binary classification, the top three most important features are "concave_points3", "perimeter3", and "concave points_1", and for multi-classification we see that for the "Adelie" and "Chinstrap" classes, the most important feature is "culmen_depth_mm", whereas for the "Gentoo" class, "flipper_length_mm" is the most important feature.

3.2 Gradient Comparison in Logistic Regression

For multi-class classification, multi-class logistic regression yielded an analytical gradient over all features and classes of 3.804071437052414e-4, and an approximate numerical gradient over all features and classes of 3.8038710752985594e-4, resulting in a sum of squared difference over all features and classes of 1.2087965765308916e-18. For binary classification, logistic regression yielded an analytical gradient over all features and classes of -2825.419019014676, and an approximate gradient over all features and classes of -2825.37682744427, resulting in a sum of squared difference over all features and classes of 5.574836649166704e-11.

3.3 Convergence Analysis of Logistic and Multi-class Logistic Regression

The convergence plots for binary and multi-classification logistic regression models are illustrated by figures 8 and 9, respectively. This plot shows the cross-entropy loss for both training and validation data over multiple iterations, providing insight into the optimization process of logistic regression. Loss is high at the early iterations but rapidly decreases as the model learns from the data, eventually stabilizing as it converges to an optimal solution. The validation loss follows a similar trend to the training loss, indicating that the model generalizes well without significant over-fitting. This behavior implies that the learning algorithm is effectively minimizing the error and reaching a steady state, ensuring reliable predictions.

3.4 ROC Curve Comparison: Logistic vs. Multiple Linear Regression

The ROC curve is depicted in Figure 1. This curve illustrates the trade-off between true positive and false positive rates across different classification thresholds, providing a comprehensive measure of model performance. While both models demonstrate near-perfect discrimination, logistic regression achieved an AUROC of 0.9995, slightly outperforming linear regression's AUROC of 0.9982. This small but notable difference suggests that logistic regression is better suited for classification tasks, as it is specifically designed to estimate probabilities and optimize decision boundaries. In contrast, linear regression, which assumes a continuous output, may not be as inherently effective for probability-based classification, even when adapted for such tasks.

3.5 Classification Accuracy for Multi-class Logistic Regression and and Multivariate Linear Regression for Multi-class classification

The multi-class logistic regression method achieved an accuracy of 95.80%, demonstrating a high level of predictive performance. In comparison, the multivariate linear regression approach for multi-class classification yielded a perfect accuracy of 100% on the test data. This suggests that the latter model was able to fit the data exceptionally well, potentially indicating the models ability to minimize the sum of squared errors for class probabilities was optimal for the classification task. However, further validation, such as cross-validation or testing on other datasets, would help confirm the generalizability of this model for multi-classification tasks.

3.6 Feature Influence in Logistic Regression

The feature importance plots for logistic regression are illustrated by Figure 5 for binary classification, and Figure 4 for multi-classification. For binary classification, the 3 most influential features were "radius2", "texture3", and "radius3". For multi-classification, the most influential feature for the Adelie and Chinstrap classes was "culmen_depth_mm", and flipper_length_mm" for the Gentoo class.

3.7 Feature-Class Relationship Heatmap in Multiclass Logistic Regression

The heatmap plot of the regression coefficients, and thus the relationship between the features and classes, is illustrated in Figure 6, and Figure 7 for multi-class logistic regression and multivariate linear regression for multi-class classification, respectively. This visualization allows us to compare the feature importance across different classes for both multi-class logistic regression and multivariate linear regression for multi-class classification. In multi-class logistic regression, the coefficients represent how strongly each feature influences the probability of a sample belonging to a particular class, often capturing non-linear relationships between features and class labels. In contrast, the multivariate linear regression is applied to one-hot encoded categorical targets, treating each class label as a continuous variable. It models feature importance based on direct linear contributions to the predicted class scores, minimizing the sum of squared errors rather than optimizing for probabilistic classification. As a result, the differences in feature importance between the two methods arise because logistic regression optimizes classification boundaries, whereas linear regression captures direct correlations between input features and output values. This distinction explains why certain features may have a greater or lesser impact depending on the modeling approach used.

3.8 Discussion on Additional Findings

As our first optional task, we experimented with the scikit-learn implementations of K Nearest Neighbours (KNN) and Decision Tree (DT) on the binary classification task. For the binary classification task, KNN achieved an accuracy of 0.95 while KNN notably underperformed, achieving an accuracy of 0.84.

Similarly, we experimented with regularized regression methods, also using libraries imported from scikit-learn. The methods chosen were Ridge regression and Lasso Regression. On the binary task Ridge yielded an accuracy of 0.97, while Lasso seriously underperformed at 0.24, signaling this method is not suited for binary classification tasks.

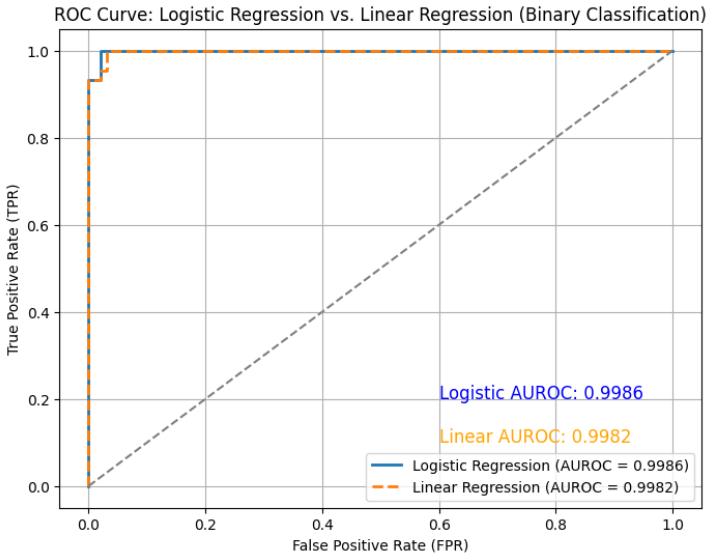

Figure 1: ROC Curve of Logistic Regression vs. Linear Regression for Binary Classification

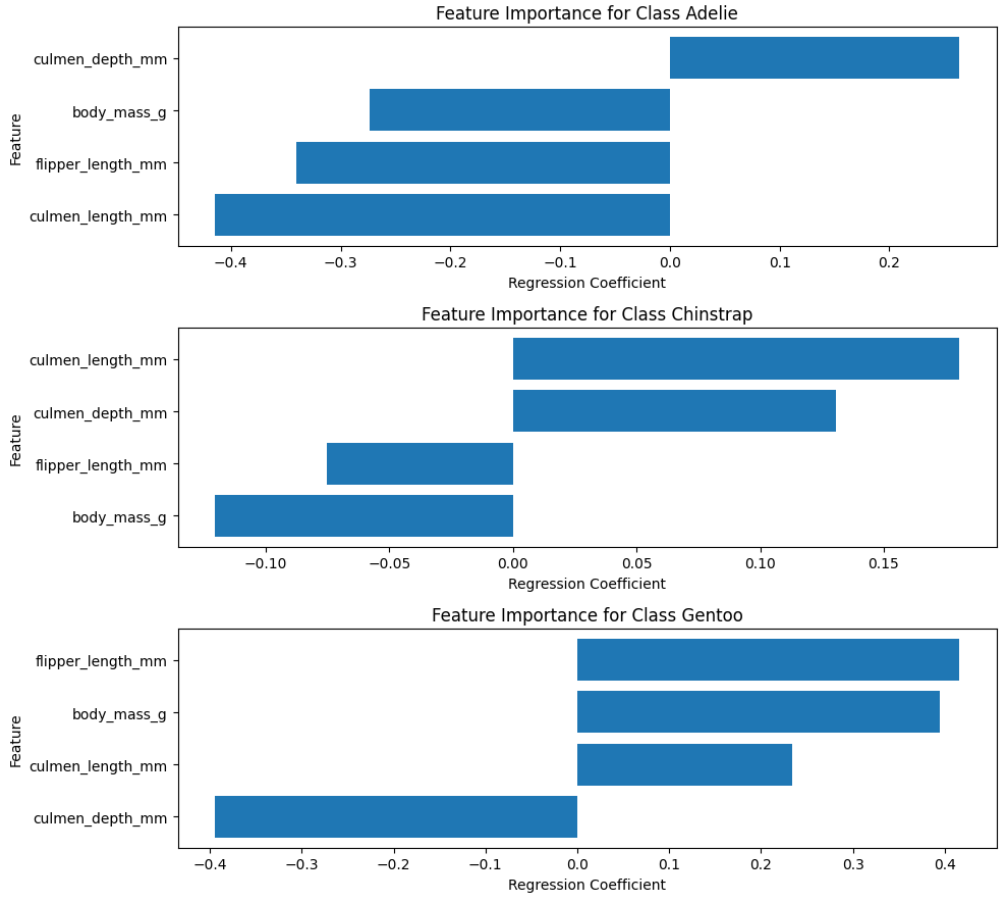

Figure 2: Feature Importance for Linear Regression (Multi-classification)

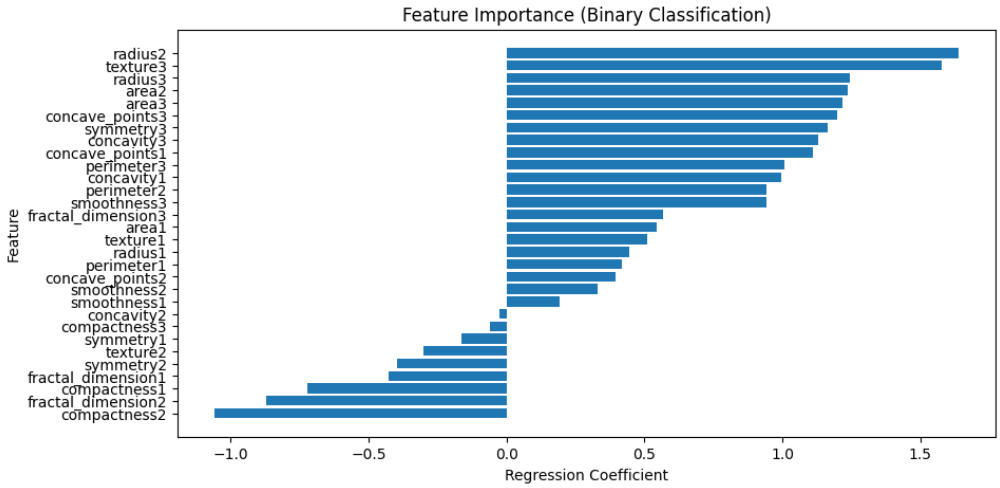

Figure 3: Feature Importance for Linear Regression (Binary Classification)

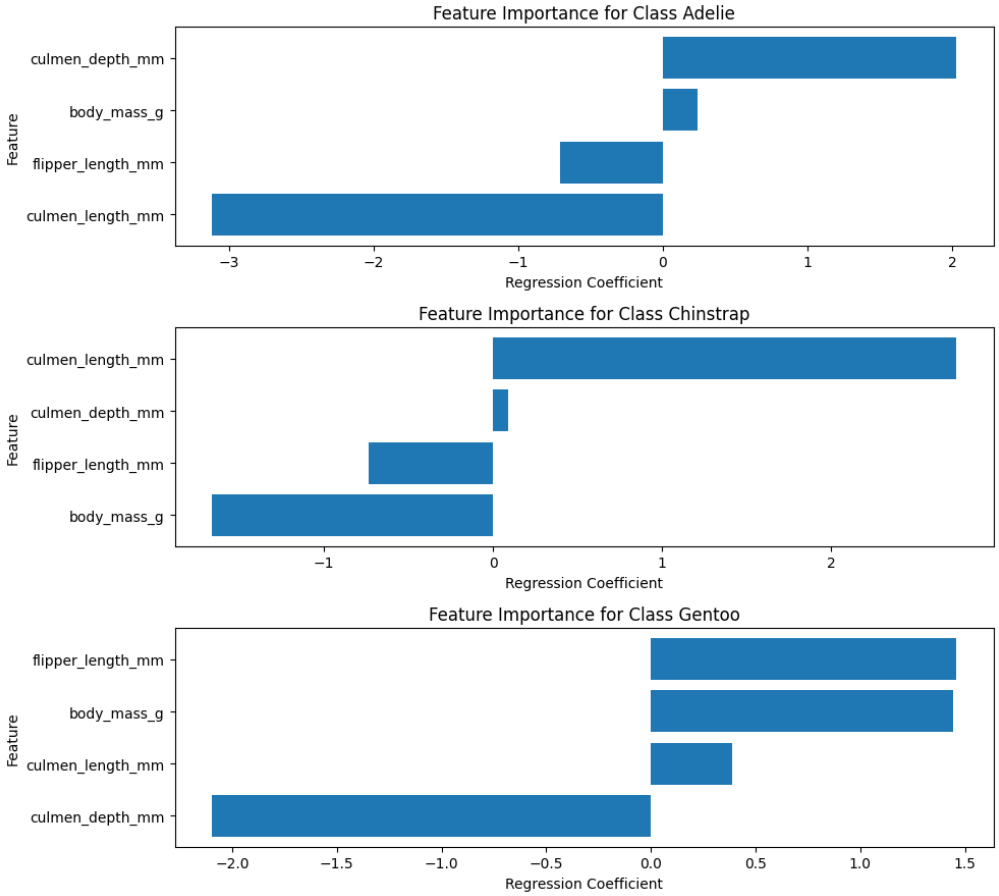

Figure 4: Feature Importance for Logistic Regression (Multi-classification)

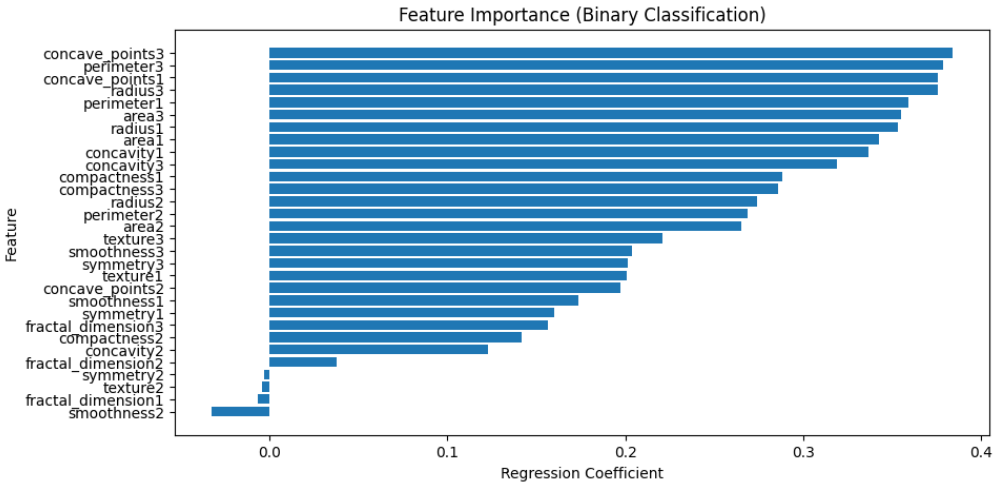

Figure 5: Feature Importance for Logistic Regression (Binary Classification)

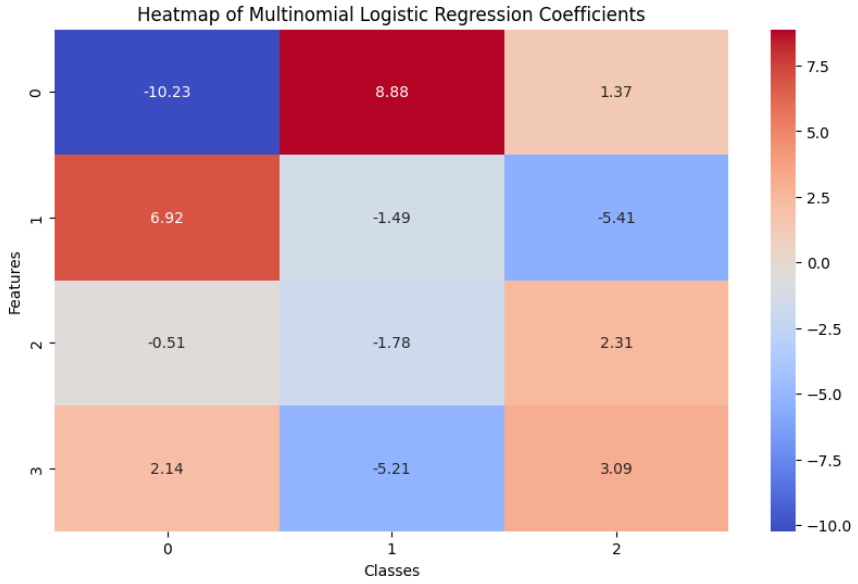

Figure 6: $D \times C$ heatmap for multi-class logistic coefficients

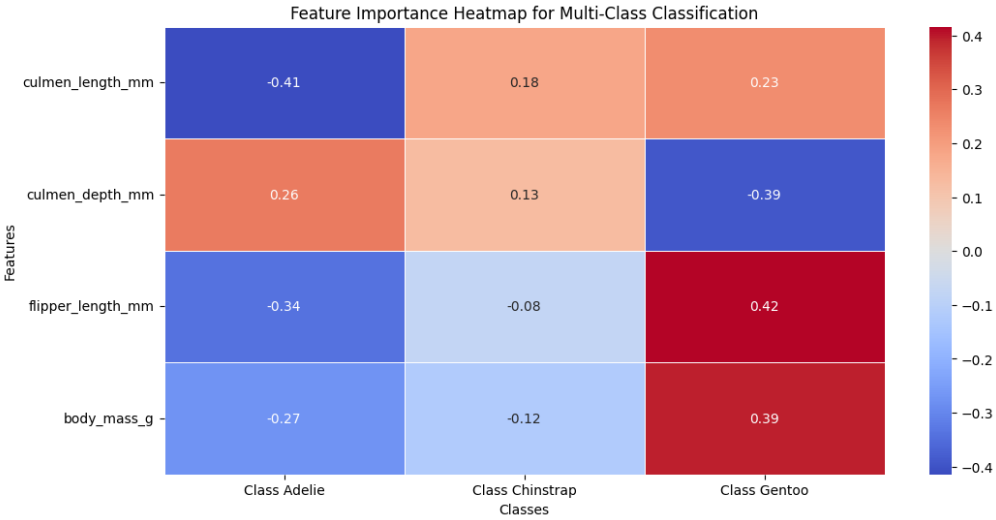

Figure 7: $D \times C$ heatmap for multi-class linear coefficients

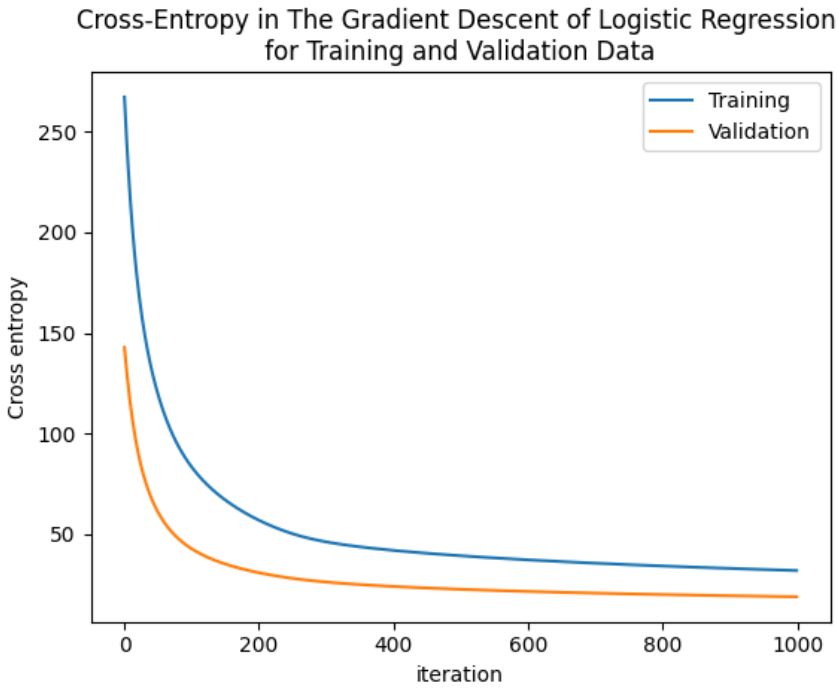

Figure 8: Convergence Curve for Logistic Regression (Binary Classification)

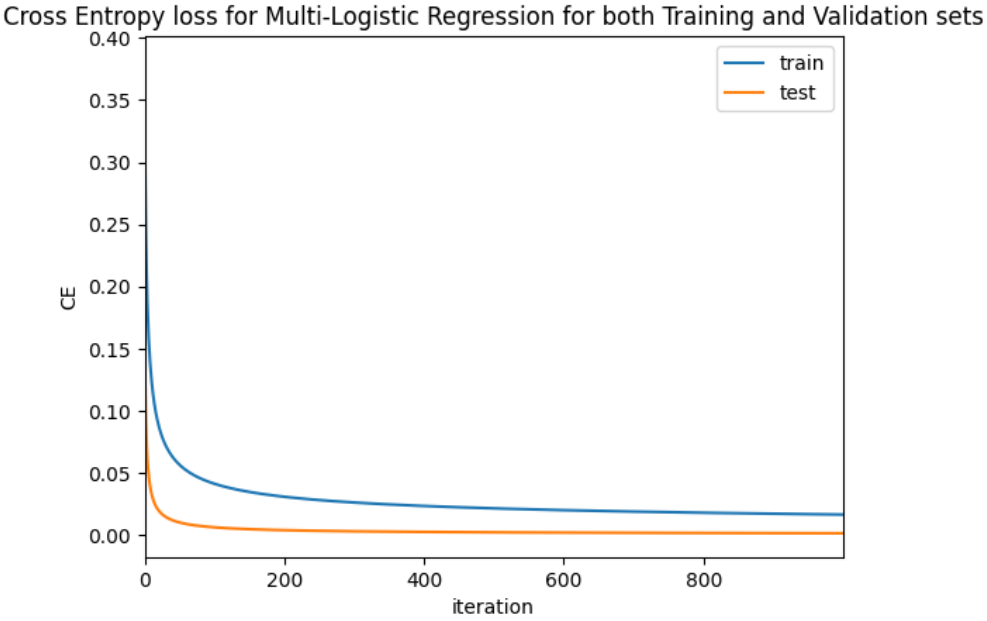

Figure 9: Convergence Curve for Logistic Regression (Multi-classification)

4 Discussion & Conclusion

Our analysis revealed that logistic regression consistently outperformed linear regression for classification tasks. Specifically, logistic regression achieved an AUROC of 0.9995, while linear regression reached only 0.9982, reaffirming that linear regression is less suitable for binary classification due to its reliance on a continuous target variable assumption. In terms of efficiency, logistic regression exhibited strong convergence behavior, demonstrating stability in its optimization process. Additionally, our multi-class logistic regression model achieved an accuracy of 95.80%, highlighting its effectiveness in handling multiple categorical outcomes.

In contrast, a surprising outcome was that the multivariate linear regression model attained 100% accuracy, raising concerns

of possible over-fitting. Since this model treats one-hot encoded categorical targets as continuous variables and minimizes squared error rather than optimizing classification boundaries, it may have memorized the training data instead of learning generalizable patterns. This issue is particularly relevant if the dataset is linearly separable or if the model lacks regularization. To assess the extent of over-fitting, evaluating performance on an independent test set or other multi-classification datasets is necessary.

A comparison of regression coefficients showed that logistic and linear regression ranked features differently, likely due to their distinct optimization objectives; linear regression minimizes squared error, while logistic regression optimizes cross-entropy loss. Feature importance analysis identified concave_points3, perimeter3, and concave_points_1 as the most relevant features for binary classification. For multiclass regression, culmen_depth_mm was significant for distinguishing Adelie and Chinstrap species, while flipper_length_mm was critical for Gentoo classification. These findings align with domain knowledge, as culmen depth and flipper length are biologically relevant characteristics for species differentiation.

Overall, our study demonstrates that logistic regression is a more suitable choice for binary classification tasks, while multi-class logistic regression effectively extends its applicability. The perfect accuracy observed in multivariate linear regression warrants further investigation, as it may indicate model over-fitting rather than true predictive power. Future work could explore additional regularization techniques or alternative classification methods to mitigate over-fitting and improve generalization. This assignment provided a comprehensive analysis of each method, offering valuable insights into their real-world applicability.

5 Statement of Contribution

Lucas A implemented linear regression, while Alina S-J handled logistic regression. Jack P-W assisted with debugging and authored the majority of the report, with Alina S-J contributing to the 'Datasets' section. All group members contributed equally to running experiments.

References

Castro, H. M., & Ferreira, J. C. (2022). Linear and logistic regression models: When to use and how to interpret them?.

Hawkins, D. M. (2004). The problem of overfitting. *Journal of Chemical Information and Computer Sciences*, 44(1), 1–12. https://doi.org/10.1021/ci0342472

Ibiyeye, T., & Ibiyeye, A. (2024). Analyzing economic impact of US trade policies and regulations on business growth using multivariate regression models. *ICONIC Research and Engineering Journals*, 8(5), 276–289.

McBride-Ellis, C. (2022, November 28). *Kaggle survey 2022: Algorithms - ML vs. DL & people* [Kaggle notebook]. Kaggle. `https://www.kaggle.com/code/carlmcbrideellis/kaggle-survey-2022-algorithms-ml-vs-dl-people`

Pandey, P. (2020). *Penguin dataset: The new iris* [Dataset]. Kaggle. Retrieved February 20, 2025, from `https://www.kaggle.com/code/parulpandey/penguin-dataset-the-new-iris/input?select=penguins\_size.csv`

Rendall, R., Castillo, I., Schmidt, A., Chin, S. T., Chiang, L. H., & Reis, M. (2019). Wide spectrum feature selection (WiSe) for regression model building. *Computers & Chemical Engineering*, 121, 99–110. https://doi.org/10.1016/j.compchemeng.2018.10.003

Su, X., Yan, X., & Tsai, C. L. (2012). Linear regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(3), 275–294. https://doi.org/10.1002/wics.1198

Tabaei, B. P., & Herman, W. H. (2002). A multivariate logistic regression equation to screen for diabetes: Development and validation. *Diabetes Care*, 25(11), 1999–2003. https://doi.org/10.2337/diacare.25.11.1999

Wolberg, W., Mangasarian, O., Street, N., & Street, W. (1993). *Breast cancer Wisconsin (diagnostic)* [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5DW2B