Assignment 2: Linear Regression and Classification

COMP 551
March 2nd, 2025

Androw Abd El Malek (261050237)
Kevin Luo (261038261)
Radu Petrescu (261051351)

Abstract

In this assignment we were tasked with developing a linear regression, a logistic regression and a multiclass logistic regression models and testing them on two distinct datasets. The first dataset chosen was a breast cancer with binary diagnosis labels, while the second was a penguins species classification dataset with 3 species predictions. Our results showed that for the binary dataset, both linear and logistic models predicted perfectly with an AUROC of 1. However for the multiclass classification data the multivariate linear regression (98.55%) outperformed the multiclass logistic regression (94.29%) in terms of accuracy on the testing set. We also analyzed feature importance through model coefficients and visualized them using bar plots and heatmaps. Overall, linear regression provided better classification performance, while logistic regression models performed slightly worse.

Introduction

The goal of this assignment is to implement linear and logistic regressions models from scratch in order to determine which model performs better in specific situations. To perform our analysis on these different models we used tabular datasets, on which we performed feature analysis and important feature extraction and later used appropriate evaluation metrics such as AUROC and hyperparameter tuning to determine performance and adaptability. The linear model was built with closed-form and gradient-descent based technique, while the logistic and multiclass regression models can only be computed using a gradient-based approach. In order to validate the numerical precision of our gradient implementation, small perturbation tests were introduced to check that the result did not differ from the expected value.

Furthermore, we compare linear regression-based classification models with logistic regression models to determine their effectiveness in different classification tasks. The first dataset chosen is the Breast Cancer Wisconsin dataset containing 569 samples with 30 continuous features with binary target variables that indicate the presence of a tumor [2]. This study was performed in 1992 and revolutionized the medical imagery field with its results as the models implemented achieved a 97% accuracy in detecting malignant tumors. The Palmer Penguins Dataset contains 342 samples with 4 continuous features with the target variable being one of 3 possible penguin species. After doing some features analysis we reduce the first set to 26 features [1]. This dataset uses a particularly low number of features as it “found in practice that beyond a certain point, the addition of further variables leads to a decrease in the classification performance” [1]. This data was used in our assignment to determine the classification performance of machine learning models, particularly regression models.

Datasets

The Breast Cancer Wisconsin Diagnostic dataset was used for binary classification [2]. This dataset contains features extracted from digitized images of fine needle aspirates (FNA) of breast masses, including attributes such as radius, texture, perimeter, area, and smoothness. To ensure data consistency, missing values were removed, and the target variable, 'Diagnosis,' was encoded as a binary variable where malignant ('M') was represented as 1 and benign ('B') as 0. To improve numerical stability and facilitate logistic regression, all features were standardized using StandardScaler, ensuring uniform feature scaling across variables. For exploratory analysis, a simple regression approach was employed to determine the importance of each feature. The computed regression coefficients revealed that radius, concavity, and perimeter were the most positively correlated with malignancy, aligning with medical knowledge that malignant tumors tend to be larger and more irregular. In contrast, smoothness exhibited a negative correlation with malignancy, as benign tumors typically have smoother contours. To optimize model efficiency, features with an absolute importance below a threshold of 0.1 were removed, leaving a refined subset of key predictors that balance interpretability with predictive power.

For multiclass classification, the Palmer Penguins dataset [1] was used to predict species based on physical characteristics. This dataset includes four continuous features—culmen length, culmen depth, flipper length, and body mass—representing morphological traits of three penguin species: Adelie, Chinstrap, and Gentoo. To enhance data quality, categorical features such as 'island' and 'sex' were removed, and missing values were dropped. The target variable 'species' was one-hot encoded, converting it into three binary columns corresponding to each species. Additionally, continuous features were standardized to have zero mean and unit variance, ensuring numerical stability for the logistic regression model. Feature importance was analyzed using regression coefficients, revealing distinct morphological traits for each species. Adelie penguins were characterized by shorter but deeper culmens, as indicated by a negative coefficient for culmen length and a positive coefficient for culmen depth. Chinstrap penguins exhibited longer and slimmer culmens, reflected in a high positive coefficient for culmen length and a smaller coefficient for culmen depth, while their lower body mass was captured by a negative regression coefficient. Gentoo penguins, on the other hand, had the longest flippers and the highest body mass, both strongly associated with high positive regression coefficients, whereas their culmen depth was relatively smaller, as shown by a negative coefficient. Since all features were deemed relevant based on this analysis, no further feature selection was performed, preserving the original four numerical features to ensure that key species characteristics were accurately represented in the model.

Results

Simple linear regression coefficients

We begin our experiments with a simplistic interpretation of our binary classification dataset using a simple linear regression model. This initial analysis helps identify key features in the dataset and further focus our attention on them when using more complex models for predictions. The plot in Fig. 1 offers valuable insights for model interpretation and optimization. The most important features, as indicated by their higher regression coefficients, are concentrated at the bottom of the chart: concave_points3, perimeter3, and concave_points1 stand out with coefficients approaching or exceeding 0.8, suggesting these features strongly influence the classification outcome. The model also places significant weight on anatomical measurements like radius3, perimeter1, area3, and radius1, all showing coefficients above 0.7. A clear pattern emerges where features related to concave points, perimeters, radii, and areas generally demonstrate higher importance than texture, symmetry, and fractal dimension features. This suggests that shape-based characteristics are more predictive for this binary classification task than textural or structural complexity measures. The least important features—texture2, fractal_dimension1, and symmetry2—have coefficients below 0.05, indicating minimal contribution to the model's decisions. These observations can help guide our attention of the most important features in the dataset, those who’s impact on predicting the presence of a tumor are highest.

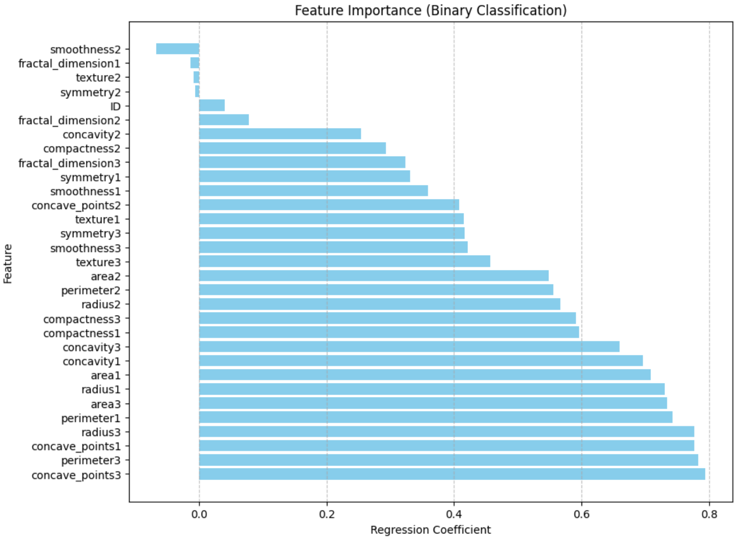

Figure 1: Simple Linear Regression Coefficient Weights for Cancer Dataset

For the second dataset, we can plot the coefficients as a heatmap for each target class of the penguins species. The heatmap (Fig. 2) represents the coefficients from a simple linear regression model, showing how each feature contributes to predicting species classification. A positive coefficient means that increasing the feature value is associated with a higher likelihood of the species classification, while a negative coefficient suggests the opposite.

The key observations from the simple linear regression model show that culmen depth has the strongest positive influence on predicting Adelie penguins (0.267609), while culmen length has the most negative impact (-0.415518). For Chinstrap penguins, the coefficient magnitudes are relatively small, with the highest being culmen depth (0.127998), suggesting they are harder to distinguish using a linear model. In the case of Gentoo penguins, flipper length (0.416785) and body mass (0.392651) have the strongest positive effects, while culmen depth (-0.395607) has a strong negative influence. These trends suggest that body size and beak dimensions play a crucial role in species classification, though a linear model may not fully capture complex relationships.

These results highlight that simple linear regression captures clear trends in how features contribute to species classification, but its effectiveness depends on how well the relationships between features and classes are truly linear.

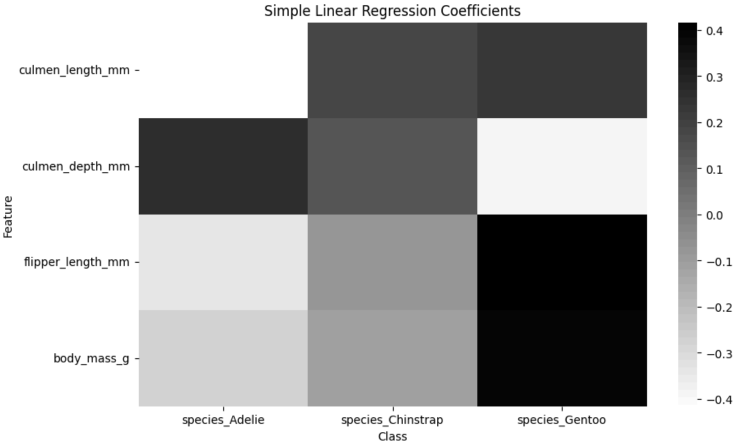

Figure 2: Simple Linear Regression Coefficient Weights for Penguins Dataset

Gradient calculations, perturbation and cross-entropy loss

A key component of the logistic and multiclass logistic regression models is the gradient descent algorithm which allows us to compute the weights of regression coefficients efficiently and quickly since no closed-form solution exists. Indeed, this iterative approach was taken for both model's *fit* methods. However, to ensure the correctness of the computed gradients in our optimization process, we performed a numerical gradient check using finite differences. Gradient verification is crucial in optimization algorithms, particularly in machine learning, where incorrect gradients can lead to slow convergence or divergence. By comparing the analytical gradients, computed via backpropagation, with numerical approximations, we can confirm whether our implementation is correct.

Our results indicate that for logistic regression, after running the optimization process for 100,000 iterations, the optimization has converged and our regression coefficients are stable. To validate the computed gradients, we compared them with their numerical counterparts. The absolute difference between the analytical and numerical gradients was found to be $\mathbf{2.131 \times 10^{-11}}$, and the relative squared error (RSE) was $\mathbf{9.547 \times 10^{-19}}$. These values indicate an extremely small discrepancy (less than $\mathbf{10^{-8}}$**)**, well within acceptable numerical precision limits. Additionally, a smooth training loss curve of the cross-entropy as can be seen in Fig. 3 indicates correct implementation. Given these results, we can confidently conclude that the analytical gradients are correctly computed, and the gradient check has successfully passed.

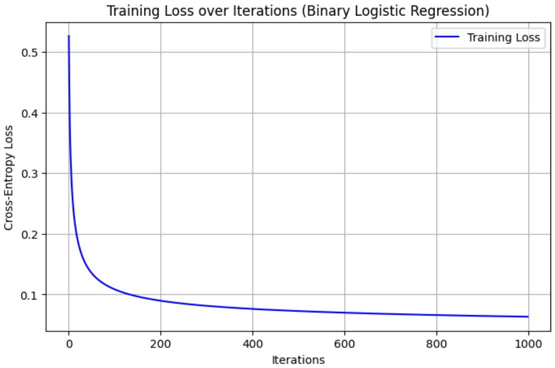

Figure 3. Training Loss of Logistic Regression

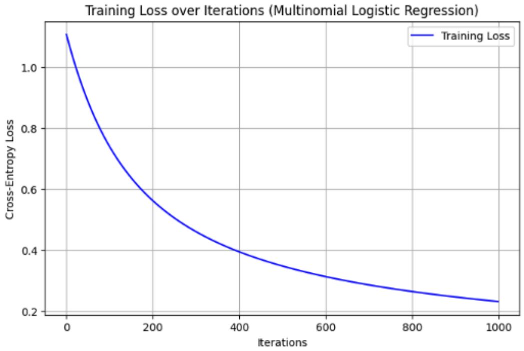

Figure 4. Training Loss of Multiclass Logistic Regression

Similarly, for the multiclass logistic regression, over 1000 iterations, the absolute difference between the analytical and numerical gradients was found to be $\mathbf{7.015 \times 10^{-11}}$. This once again indicates an extremely small discrepancy (less than $\mathbf{10^{-8}}$**)**, well within acceptable numerical precision limits. Then, the training loss curve for this model is still smooth, but having the initial loss larger than the other model, shown by the forest few iterations having CE loss > 1 (Fig. 4), while the logistic model started around 0.5 (Fig 3.). The steepness of the curve is also lesser since less iterations are used.

ROC of linear and logistic regression models on binary classification

Having verified the correctness of our models, we proceeded to test them on our datasets. We first used the binary classification dataset for breast cancer, applying both logistic regression and multiple linear regression models to classify the data. Using a training set (80%), validation set (10%), and testing set (10%), we selected the best model by evaluating different hyperparameters on the validation set. The best-performing model on the validation set was then used to evaluate the test set AUROC. To ensure the models converged to an optimal solution, we trained them over multiple epochs. An epoch represents one full pass through the training data during optimization. For logistic regression, we implemented an iterative training process where the model updated its weights over 100 epochs while monitoring the AUROC on the validation set. If the validation AUROC improved, the model's parameters were saved, and the patience counter was reset. However, if no improvement was observed for 10 consecutive epochs, early stopping was triggered to prevent overfitting. The final model was then evaluated on the test set, achieving a strong AUROC score. Finally, the test set accuracy was computed, confirming the effectiveness of our approach. The ROC curve was plotted based on the predictions, providing a visual representation of each model's performance.

The ROC in figure 5 curve compares the performance of linear regression, logistic regression, and a random classifier in a binary classification task. The area under the ROC curve (AUROC) values indicate how well each model can distinguish between the two classes. Linear regression, despite not being intended for classification, achieves an AUROC of 1, which is surprisingly high and suggests that it can effectively separate classes in this particular dataset. Logistic regression, designed specifically for classification, performs the same with an AUROC of 1. The random classifier, serving as a baseline, has an AUROC of 0.5, which represents purely random guessing. The close performance of logistic and linear regression suggests that the decision boundary in this dataset may be nearly linear.

A perfect AUROC of 1 suggests that the dataset is highly linearly separable, meaning the features provide a clear distinction between the two classes. Both logistic and linear regression can perfectly classify the data because the decision boundary aligns well with the feature distributions. This result is highly dependent on the specific training and testing sets—modifying them can alter AUROC. Based on these results, both models perform equally in the binary classification task on the breast cancer dataset.

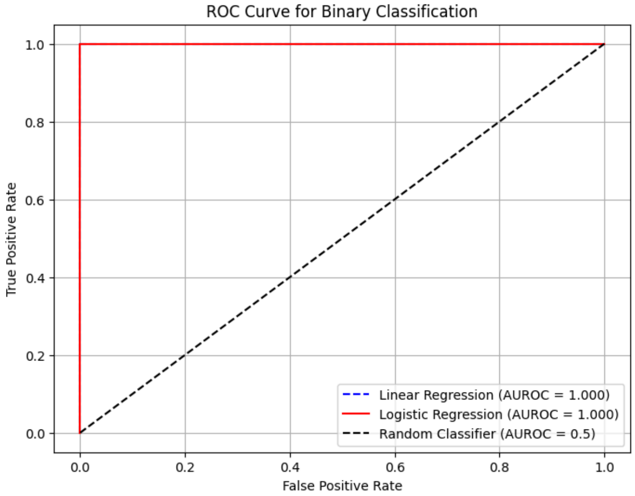

Figure 5. ROC curve for logistic regression and multiple linear regression on the Breast Cancer Dataset

Classification accuracy of multiclass logistic and multiple regression on multiclass classification

Moving on to the multiclass classification dataset on the penguins, we took a similar approach to test the multivariate linear regression model and the multiclass logistic regression model. A similar process to logistic regression was applied to the multiclass logistic regression model, where accuracy was used as the evaluation metric instead of AUROC. The model was trained over multiple epochs, with each epoch representing a full pass through the training data. During each epoch, the model updated its weights using gradient descent, and its performance on the validation set was monitored. If the validation accuracy improved, the model's parameters were saved; otherwise, a patience counter was incremented. Training stopped early if no improvement was observed for a set number of consecutive epochs, ensuring that the model did not overfit to the training data. Then, for the multiclass logistic regression model, we used the softmax function to convert raw model outputs (logits) into probability distributions over the classes. This allowed the model to assign probabilities to each class and make predictions based on the highest probability. Softmax ensured that the predicted class probabilities summed to one, making it an effective method for multiclass classification.

After training, we evaluated both models on the test data. The multiclass logistic regression model achieved a test accuracy of 94.29%, while the multivariate linear regression model outperformed it with a test accuracy of 98.55%. These results demonstrate that both models effectively classified the penguin dataset, with multivariate linear regression yielding the highest accuracy. While accuracy is an important metric for understanding model performance, it is also valuable to explore how individual features contribute to these predictions. This can provide deeper insights into the models' decision-making processes and guide potential improvements.

Feature coefficients of models for binary classification

Moving back to the binary classification dataset we perform a feature analysis in order to better understand the influence of each feature on the predictions. We now turn to a visualization of the coefficients from the logistic regression model on the binary classification dataset (Cancer). As can be seen in the graph (Fig. 6), the logistic regression plot shows the magnitude and direction (positive or negative) of each feature's contribution to the classification decision. Features like concave_points 3 and perimeter2 have the highest positive coefficients, indicating that they play a crucial role in predicting the presence of a tumor. On the other hand, features such as concavity2 and compactness2 have very small coefficients, meaning they contribute very little to the prediction. Most regression coefficients are very similar suggesting that most features play a more-or-less significant role in the prediction of the cancer.

In the multiple linear regression model, the coefficient values vary significantly, with some features having much larger impacts than others. The radius3 and perimeter1 features show the largest positive coefficients, implying a strong positive influence on the predicted outcome. Other features, such as area3 and radius1, also have larger negative weights, but their coefficients are relatively smaller. The disparity in coefficient magnitudes suggests that some features are much more influential than others in predicting the target variable, which could indicate potential multicollinearity or scaling effects. Importantly, in this model some weights are negative.

The last plot is the same as in Fig. 1 but added for comparison. Overall, this visualization highlights the different ways these regression models assign importance to features. Multiple linear regression appears to be more sensitive to feature magnitudes, while logistic regression maintains a more balanced distribution of feature importance.

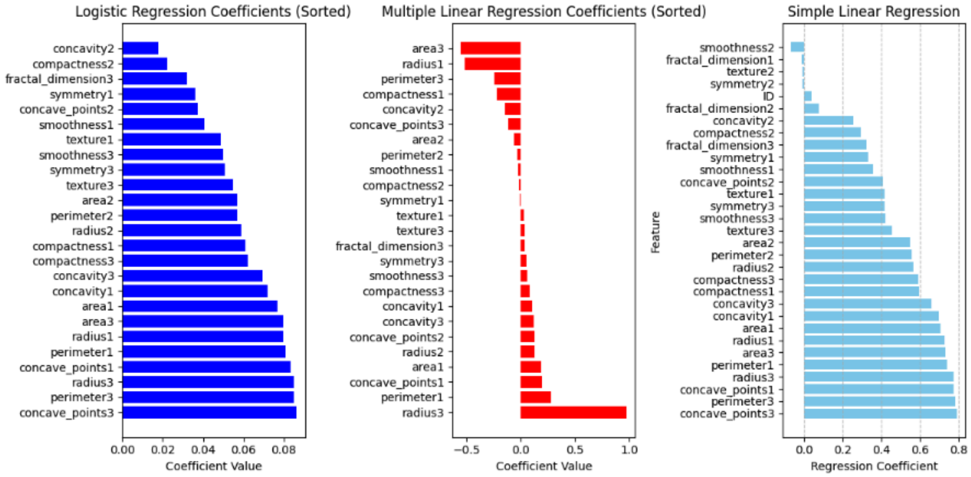

Figure 6: Weights of coefficients for logistic, multiple linear and simple linear regression models

Heatmap relationship for multiclass logistic classification

To visualize the feature importance in the multiclass classification dataset, we can use a heatmap. The heatmap (Fig. 7) visualizes the learned coefficients from a multiclass logistic regression model, illustrating how different features contribute to classifying Adelie, Chinstrap, and Gentoo penguins. Each cell represents the weight assigned to a feature for a specific class, with darker shades indicating higher values. Notably, culmen depth has the highest coefficient for Adelie penguins (0.257981), making it the most influential feature for their classification. In contrast, Chinstrap and Gentoo classifications are more influenced by culmen length, with coefficients of 0.253330 and 0.253164, respectively. Flipper length and body mass show relatively uniform contributions across all species, with coefficients ranging between 0.247 and 0.253, suggesting that these features are less discriminative. Overall, culmen depth is the strongest predictor for Adelie penguins, while culmen length and body mass play a more significant role in distinguishing Chinstrap and Gentoo species.

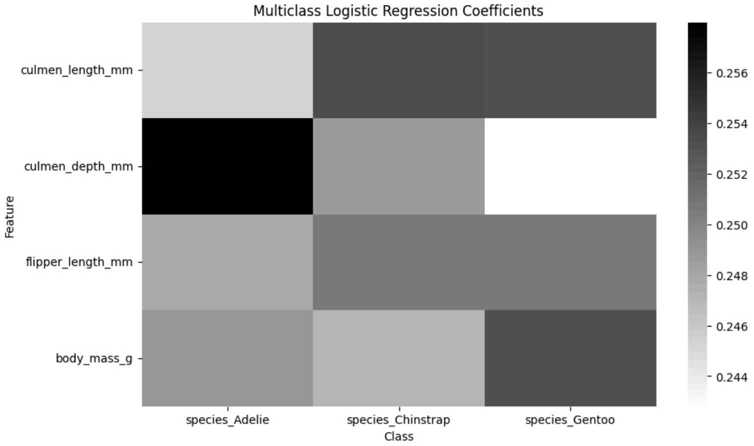

Figure 7: Multiclass logistic regression relationship between features and classes

Ridge and Lasso regression models for feature importance

In order to confirm our findings, particularly those of our features weights coefficient importance in determining the final prediction, we decided to implement Ridge and Lasso regression models to see how they would calculate the coefficients.

The Ridge regression model maintains a diverse set of features with both positive and negative coefficients. Radius3 showed the highest positive influence with a coefficient approaching 0.8, followed by concave_points1 at approximately 0.2. Notable negative coefficients appear for area3, compactness1, and perimeter3, suggesting these features inversely relate to the target variable. Ridge regression preserved most coefficients in the model rather than eliminating them entirely, which is characteristic of L2 regularization. The model achieved consistent 99.03% accuracy across all alpha values (0.2-1.0), demonstrating remarkable stability regardless of regularization strength.

The Lasso regression model drastically reduced the feature space, retaining only four features with non-zero coefficients while zeroing out the rest—a hallmark of L1 regularization's feature selection capability. Concave_points3 dominates with a coefficient of approximately 0.16, followed by radius3 (0.12), concave_points1 (0.03), and texture3 (0.02). However, this aggressive feature elimination came at a significant performance cost, with accuracy dropping from 91.30% at alpha=0.2 to just 66.67% at alpha values of 0.4 and above. This demonstrates how increasing regularization strength in Lasso leads to more features being eliminated and potentially important predictive information being lost.

Ridge maintains model complexity with stable performance across regularization strengths, while Lasso aggressively performs feature selection at the cost of accuracy as regularization increases. For this particular dataset, Ridge's approach of keeping all features with smaller coefficients proves significantly more effective than Lasso's sparse representation, suggesting complex interrelationships between features that Ridge better preserves. Both models kept similar important features such as radius3 in the top, confirming our initial calculations with linear and logistic regression models.

Comparing with KNN and DT

The final experiment compares the Linear and logistic models on the binary classification dataset (cancer) to KNN and DT models. As can be seen in figure 8, when comparing the results between linear and logistic regression and KNN or Decision Tree models, the prior perform better with the AUROC = 1, while KNN achieves second best of 0.9840 and DT is last with 0.9252. These results indicate that the regression classifiers are better at predicting on this type of data given its target values being binary.

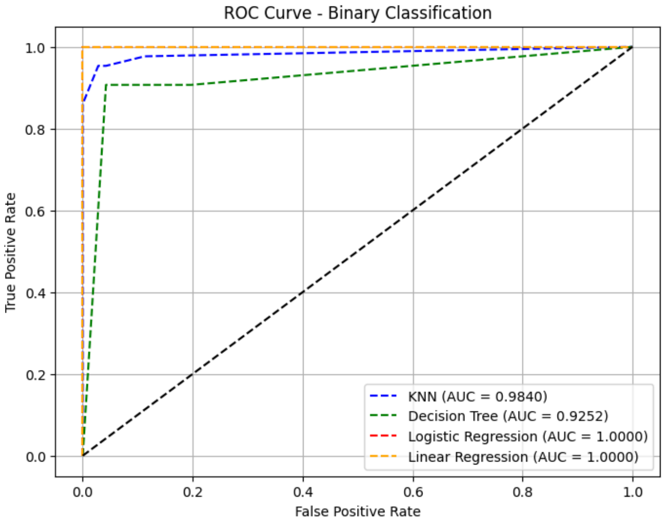

Figure 8: ROC curve of all models for binary classification dataset

However, on a multiclass classification dataset, the KNN and Linear regression performed best (98.55%), while DT came in second (95.65%) and Multiclass Logistic regression performed worst (91.43%).

Discussion and Conclusion

Our experiments demonstrate that both linear and logistic regression models can effectively perform classification tasks on structured datasets, with linear regression surprisingly matching or outperforming logistic regression in both the binary cancer dataset (both achieved perfect AUROC of 1) and the multiclass penguins dataset (98.55% vs 94.29% accuracy). Linear regression proved more computationally efficient due to its closed-form solution capability, while logistic regression required iterative gradient descent optimization. Despite logistic regression's cross-entropy loss being theoretically more appropriate for classification by directly modeling probability distributions, linear regression's squared error loss performed remarkably well when datasets exhibited strong linear separability between classes. Feature importance analysis revealed that while both models identified similar key features (concave points, radius, and perimeter measurements in the cancer dataset), linear regression showed more pronounced coefficient magnitudes with both positive and negative values, whereas logistic regression maintained more balanced coefficient distributions. The effectiveness of linear regression for classification in these cases suggests that when dataset classes are highly separable with clear linear boundaries, simpler models may outperform those specifically designed for classification tasks.

Statement of Contribution

Each team member contributed to the coding and writing part rather equally. Pair programming sessions were held where all members contributed to ensure full understanding of the models and implementation before going on to produce results (graphs) and writing the report.

References

[1] Horst, A. M., Hill, A. P., & Gorman, K. B. (2020). Palmer Penguins: A dataset for data science & machine learning. Retrieved from https://allisonhorst.github.io/palmerpenguins/

[2] Wolberg, W. H., & Mangasarian, O. L. (1995). Machine learning applied to breast cancer diagnosis and prognosis. *Cancer Letters*, 77(1), 163-171.