RIGERS BUSHI (BUS21554437)

# COURSEWORK REPORT

## Multi-Class Classification of Iris Flowers Using Logistic Regression

## 1. Introduction

Iris blossom classification into three species is one of the core tasks in supervised machine learning. This study explores the use of logistic regression for multi-class classification using the well-known Iris dataset. The primary goal is to develop a model that predicts a flower's species based on four attributes: petal length, petal width, sepal length, and sepal width. The model is evaluated using metrics such as area under the curve, confusion matrix, and accuracy.

## 2. Logistic Regression Algorithm

A probabilistic linear classifier that works well for both binary and multi-class classification applications is logistic regression. In this research, we expand logistic regression for multi-class classification using the one-vs-all technique. The one-vs-all method predicts probabilities that are added up to determine the more likely class by fitting a binary classifier for every class.
Given the linear separability of the dataset and the requirement for findings that are interpretable, logistic regression was chosen. Additionally, it is a dependable approach for small datasets like Iris due to its efficiency and simplicity.

## 3.Experiment

The Iris dataset consists of 150 samples distributed equally across three classes. Each sample is described by four numerical features:

- **Sepal Length**
- **Sepal Width**
- **Petal Length**
- **Petal Width**

**Pre-process**

1. The Id column, a unique identifier, is dropped as it holds no predictive value.
2. The target variable (Species) is encoded into numerical values using LabelEncoder.

3. Feature values are standardized to ensure uniformity in the scale of inputs.

**Data Splitting**
The dataset is divided into training and test sets with a 70-30 ratio, ensuring the class distribution remains balanced using stratified sampling.

**Model Configuration**
The logistic regression model is initialized with the following parameters:

- Multi-class strategy: one-vs-all
- Maximum iterations: 200
- Solver: lbfgs

# 4. Evaluation Metrics

To assess model performance, the following metrics are used:

1. **Accuracy**: The proportion of correctly classified samples.
2. **Confusion Matrix**: A matrix summarizing true positives, false positives, true negatives, and false negatives.
3. **ROC Curve & AUC**: The ROC curve illustrates the trade-off between sensitivity and specificity for each class, and the AUC quantifies the classifier's overall ability to distinguish between classes.

# 5.Experiments and Parameter Tuning

**Experiments Conducted**
**Data Splitting Ratios**: 80-20, 70-30, and 60-40 splits were tested to evaluate model generalization.
**Parameter Tuning**: The maximum iterations were varied between 100, 200, and 500 to ensure convergence.
**Feature Selection**: The model was tested with subsets of to analyze their individual contributions.
**Results**

| Split Ratio | Accuracy | AUC (Class 0) | AUC (Class 1) | AUC (Class 2) |
|---|---|---|---|---|
| 80-20 | 93% | 1.00 | 0.95 | 0.93 |
| 70-30 | 96% | 1.00 | 0.96 | 0.96 |

| Split Ratio | Accuracy | AUC (Class 0) | AUC (Class 1) | AUC (Class 2) |
|---|---|---|---|---|
| 60-40 | 88% | 0.99 | 0.91 | 0.92 |

Accuracy: 0.93

Classification Report:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 1.00 | 1.00 | 1.00 | 10 |
| 1 | 0.90 | 0.90 | 0.90 | 10 |
| 2 | 0.90 | 0.90 | 0.90 | 10 |
| accuracy | | | 0.93 | 30 |
| macro avg | 0.93 | 0.93 | 0.93 | 30 |
| weighted avg | 0.93 | 0.93 | 0.93 | 30 |

Accuracy: 0.96

Classification Report:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 1.00 | 1.00 | 1.00 | 15 |
| 1 | 0.93 | 0.93 | 0.93 | 15 |
| 2 | 0.93 | 0.93 | 0.93 | 15 |
| accuracy | | | 0.96 | 45 |
| macro avg | 0.96 | 0.96 | 0.96 | 45 |
| weighted avg | 0.96 | 0.96 | 0.96 | 45 |

# 6. Analysis and Discussion

On an 80-20 split, the logistic regression model's maximum accuracy was 93%, and its AUC values were consistently high for every class. When only petal measurements were employed, the accuracy increased, indicating that petal traits were more predictive than sepal features. With an AUC of 1.00, the ROC curves showed that Iris-setosa had nearly complete separability, although Iris-versicolor and Iris-virginica displayed minor overlaps because of their similar biology. Convergence was guaranteed without overfitting by raising max_iter. However, with smaller training sets, the model's performance somewhat declined, underscoring the need of having enough training data.

# 7. Conclusion

The logistic regression model demonstrated excellent performance on the Iris dataset, achieving high accuracy and AUC values. The results highlight the simplicity and efficiency of logistic regression for multi-class classification tasks. Future work could involve exploring more complex models or testing the model's robustness on noisy darta.

# 8.References


https://scikit-learn.org/stable/auto_examples/model_selection/plot_roc.html
https://stackoverflow.com/questions/61526287/how-to-add-correct-labels-for-seaborn-confusion-matrix
https://stackoverflow.com/questions/42822318/learning-curve-error