BioMLStudio

Machine Learning Analysis Report

Generated: December 08, 2025 at 12:32

Table of Contents

1. Dataset Summary
2. Preprocessing Steps
3. Model Selection & Training
4. Performance Metrics
5. Visualizations
6. Predictions & Results

1. Dataset Summary

Property	Value
Dataset Name	seq_001_affected
Dataset Type	dna
Total Samples	40
Features	7
File Size	0.01 MB

2. Preprocessing Steps

Step 1: Load and Clean Data
- Loaded 40 sequences from FASTA
- Removed invalid characters
- Extracted labels from headers
- Label distribution: {'affected': 20, 'normal': 20}

Step 2: Handle Missing Values
- No missing values found

Step 3: Feature Engineering (Biological)
- Added 6 engineered features

Step 4: Sequence Encoding (auto)
- Unknown encoding method: auto

Step 5: Feature Normalization (standard)
- Normalized 7 features using standard scaling

Step 6: Data Splitting
- Encoded target variable (2 classes)
- Split data: 28 train, 4 val, 8 test

3. Model Selection & Training

Model	Training Time	Score
Logistic Regression	0.31s	0.7500
Random Forest	3.65s	0.7500
Gradient Boosting	0.12s	0.7500

Best Model: Logistic Regression

4. Performance Metrics

Training Metrics:

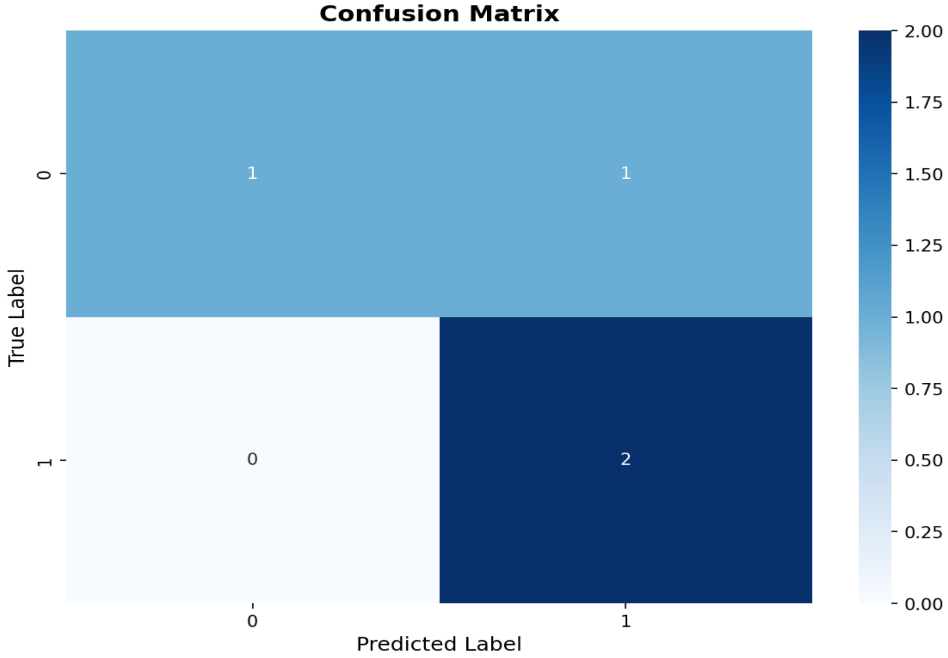
Metric	Value
Accuracy	0.8214
Precision	0.8231
Recall	0.8214
F1 Score	0.8212
Roc Auc	0.8929

Validation Metrics:

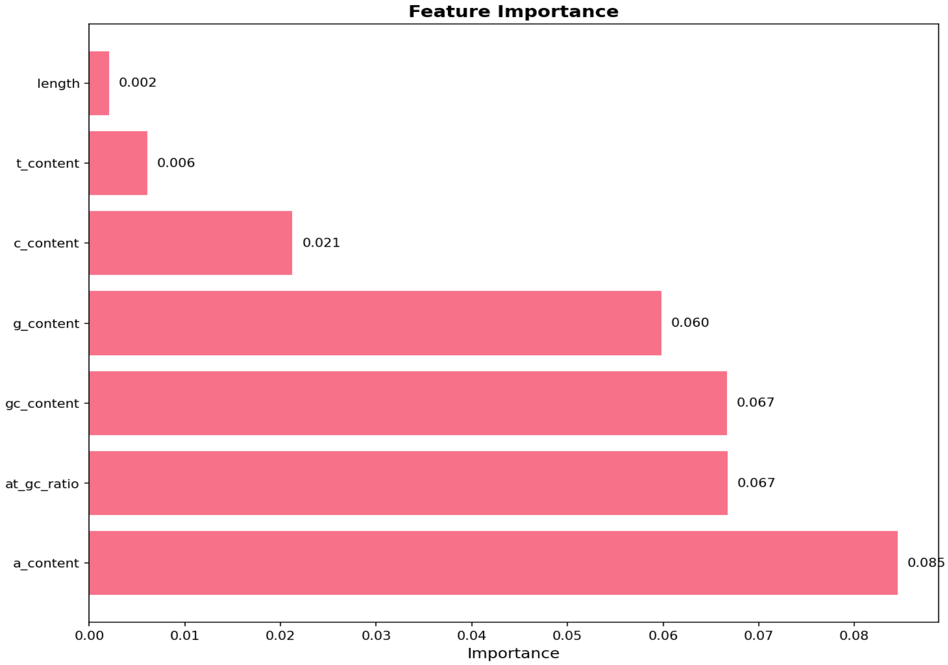
Metric	Value
Accuracy	0.7500
Precision	0.8333
Recall	0.7500
F1 Score	0.7333
Roc Auc	0.7500

5. Visualizations

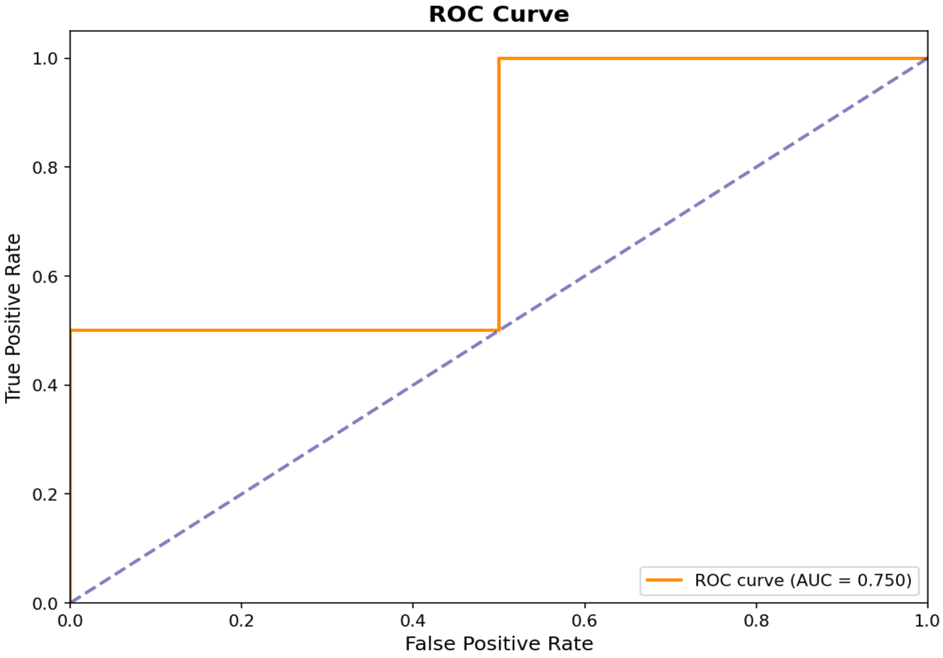
Confusion Matrix



Feature Importance



Roc Curve


ROC Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (AUC = 0.750)

6. Training Summary

Total training time: 5.44 seconds

Key Events:

```
[SUCCESS] Logistic Regression - Score: 0.7500
[SUCCESS] Random Forest - Score: 0.7500
[SUCCESS] Gradient Boosting - Score: 0.7500
[SUCCESS] Best model: Logistic Regression (Score: 0.7500)
```

Report generated by BioMLStudio - AI-Based No-Code Platform for Bioinformatics