BioMLStudio

Machine Learning Analysis Report

Generated: December 03, 2025 at 20:03

Table of Contents

1. Dataset Summary

Property	Value
Dataset Name	seq_001_unaffected.fasta
Dataset Type	dna
Total Samples	120
Features	71
File Size	0.01 MB

2. Preprocessing Steps

Step 1: Load and Clean Data
• Loaded 120 sequences from FASTA
• Removed invalid characters
• Extracted labels from headers
• Label distribution: {'affected': 120}

Step 2: Handle Missing Values
• No missing values found

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

Step 4: Sequence Encoding (kmer)
• Applied k-mer encoding (k=3)

Step 5: Feature Normalization (standard)
• Normalized 71 features using standard scaling

Step 6: Data Splitting
• Encoded target variable (1 classes)
• Split data: 84 train, 12 val, 24 test

3. Model Selection & Training

Model	Training Time	Score
Random Forest	0.13s	1.0000

Best Model: Random Forest

4. Performance Metrics

Training Metrics:

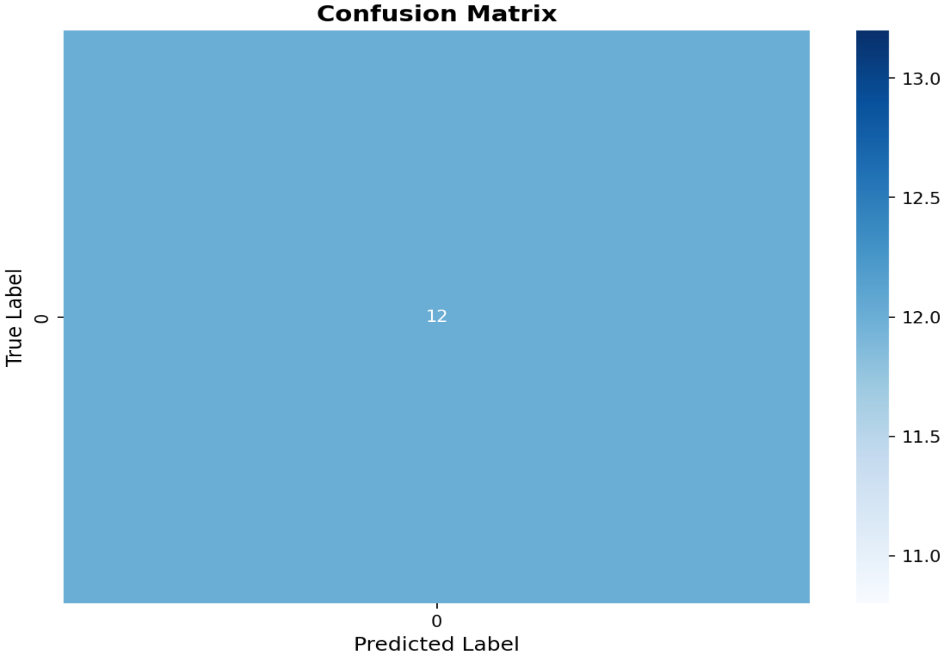
Metric	Value
Accuracy	1.0000
Precision	1.0000
Recall	1.0000
F1 Score	1.0000

Validation Metrics:

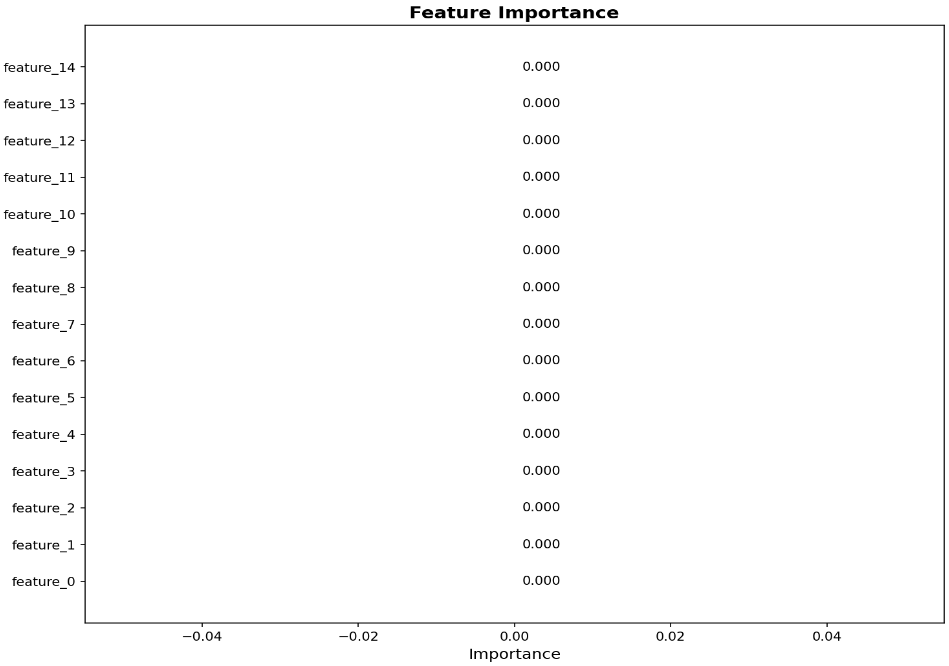
Metric	Value
Accuracy	1.0000
Precision	1.0000
Recall	1.0000
F1 Score	1.0000

5. Visualizations

Confusion Matrix



Feature Importance



6. Training Summary

Total training time: 0.62 seconds

Key Events:

```
[ERROR] Error training Logistic Regression: This solver needs samples of at least 2 classes in the data, but the data contains only one class: 0
[SUCCESS] Random Forest - Score: 1.0000
[ERROR] Error training Gradient Boosting: y contains 1 class after sample_weight trimmed classes with zero weights, while a minimum of 2 classes are required.
[SUCCESS] Best model: Random Forest (Score: 1.0000)
```

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