BioMLStudio

Machine Learning Analysis Report

Generated: December 07, 2025 at 11:58

Table of Contents

1. Dataset Summary

Property	Value
Dataset Name	SEQ001_NORMAL_HUMAN_GENE.fasta
Dataset Type	dna
Total Samples	7
Features	71
File Size	0.00 MB

2. Preprocessing Steps

Step 1: Load and Clean Data
• Loaded 7 sequences from FASTA
• Removed invalid characters
• Extracted labels from headers
• Label distribution: {'GENE': 3, 'REGION': 2, 'TYPE1': 1, 'TYPE2': 1}

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 (4 classes)
• Split data: 4 train, 1 val, 2 test

3. Model Selection & Training

Model	Training Time	Score
Random Forest	0.37s	0.0000
Logistic Regression	0.04s	0.0000

Best Model: Random Forest

4. Performance Metrics

Training Metrics:

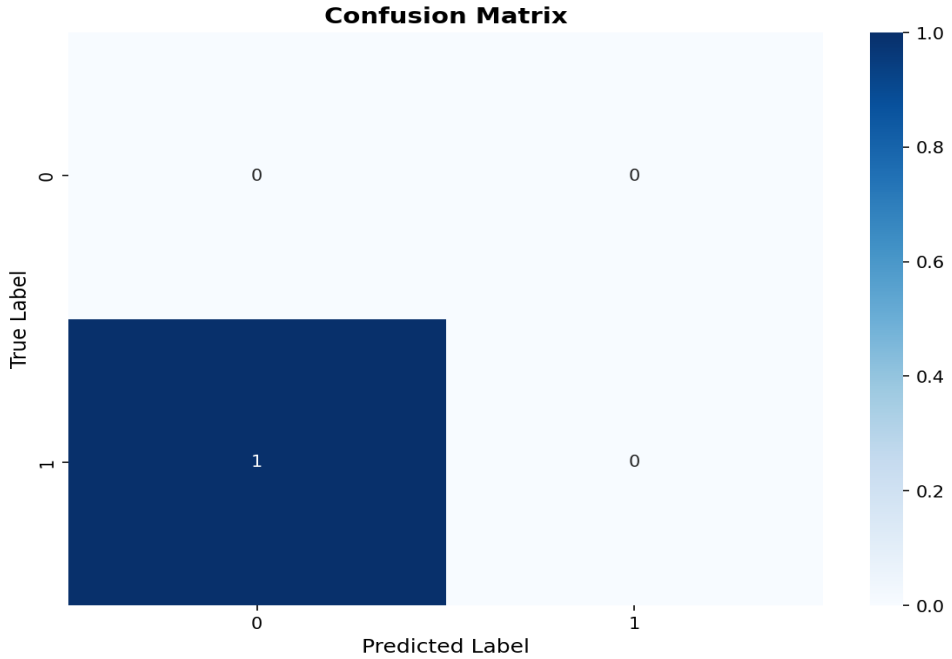
Metric	Value
Accuracy	1.0000
Precision	1.0000
Recall	1.0000
F1 Score	1.0000

Validation Metrics:

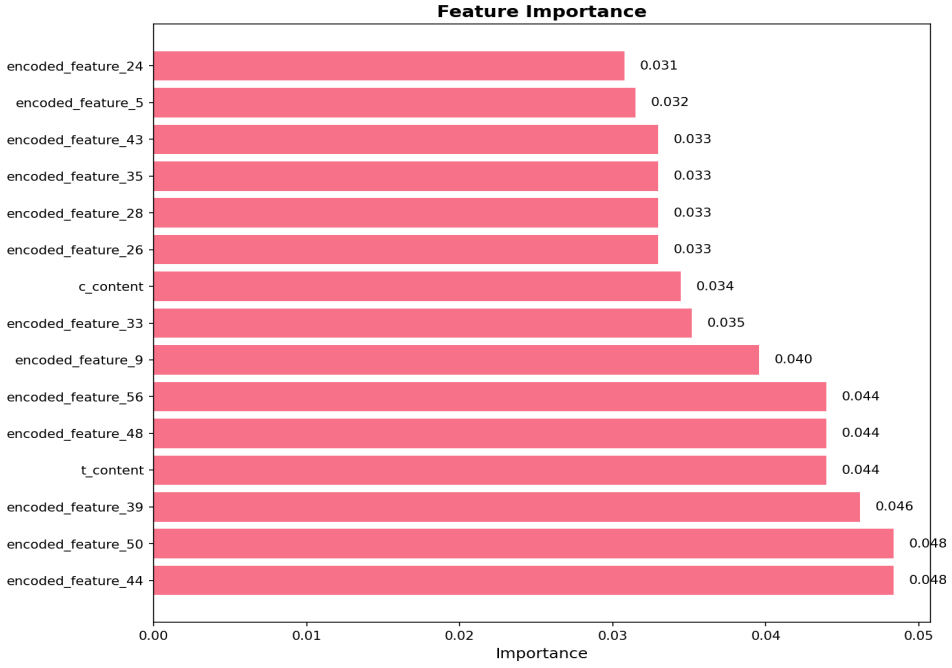
Metric	Value
Accuracy	0.0000
Precision	0.0000
Recall	0.0000
F1 Score	0.0000

5. Visualizations

Confusion Matrix



Feature Importance



6. Training Summary

Total training time: 1.74 seconds

Key Events:
[SUCCESS] Random Forest - Score: 0.0000
[SUCCESS] Logistic Regression - Score: 0.0000
[ERROR] Error training XGBoost: Invalid classes inferred from unique values of `y`. Expected: [0 1 2], got [0 1 3]
[SUCCESS] Best model: Random Forest (Score: 0.0000)

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