

# BioMLStudio

## Machine Learning Analysis Report

Generated: December 04, 2025 at 23:11

## 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.fasta |
| Dataset Type  | dna                    |
| Total Samples | 40                     |
| Features      | 71                     |
| 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 (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 (2 classes)
- Split data: 28 train, 4 val, 8 test

### 3. Model Selection & Training

| Model               | Training Time | Score  |
|---------------------|---------------|--------|
| Logistic Regression | 4.07s         | 1.0000 |
| Random Forest       | 0.40s         | 1.0000 |
| Gradient Boosting   | 0.21s         | 1.0000 |

**Best Model:** Logistic Regression

## 4. Performance Metrics

### Training Metrics:

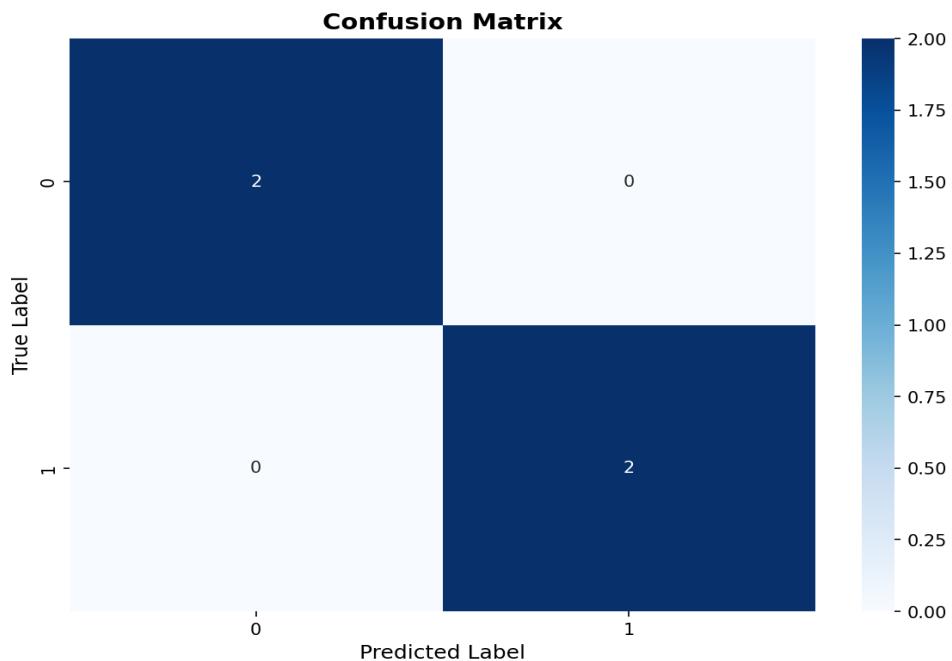
| Metric    | Value  |
|-----------|--------|
| Accuracy  | 1.0000 |
| Precision | 1.0000 |
| Recall    | 1.0000 |
| F1 Score  | 1.0000 |
| Roc Auc   | 1.0000 |

### Validation Metrics:

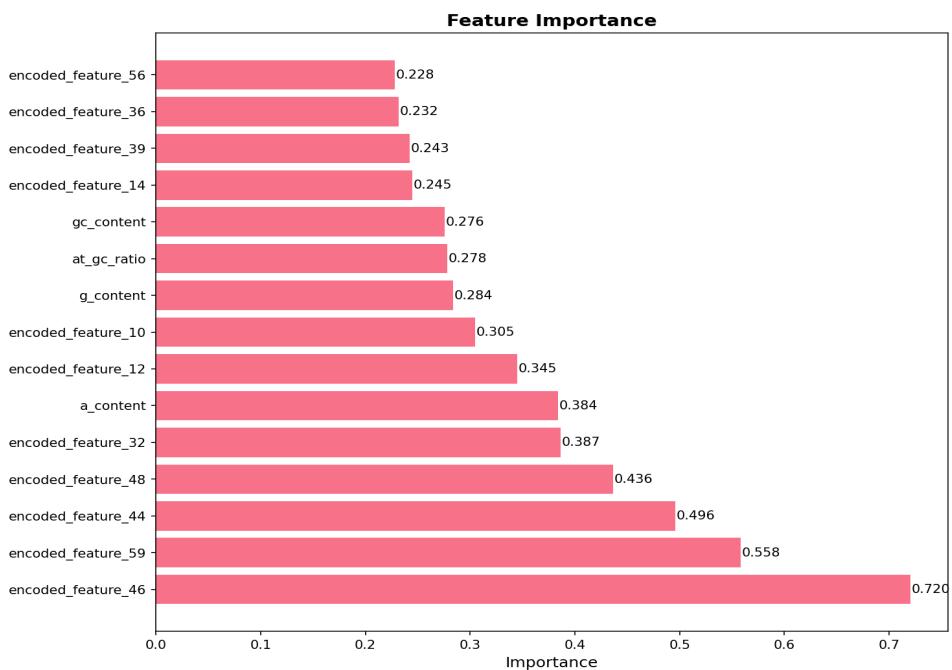
| Metric    | Value  |
|-----------|--------|
| Accuracy  | 1.0000 |
| Precision | 1.0000 |
| Recall    | 1.0000 |
| F1 Score  | 1.0000 |
| Roc Auc   | 1.0000 |

## 5. Visualizations

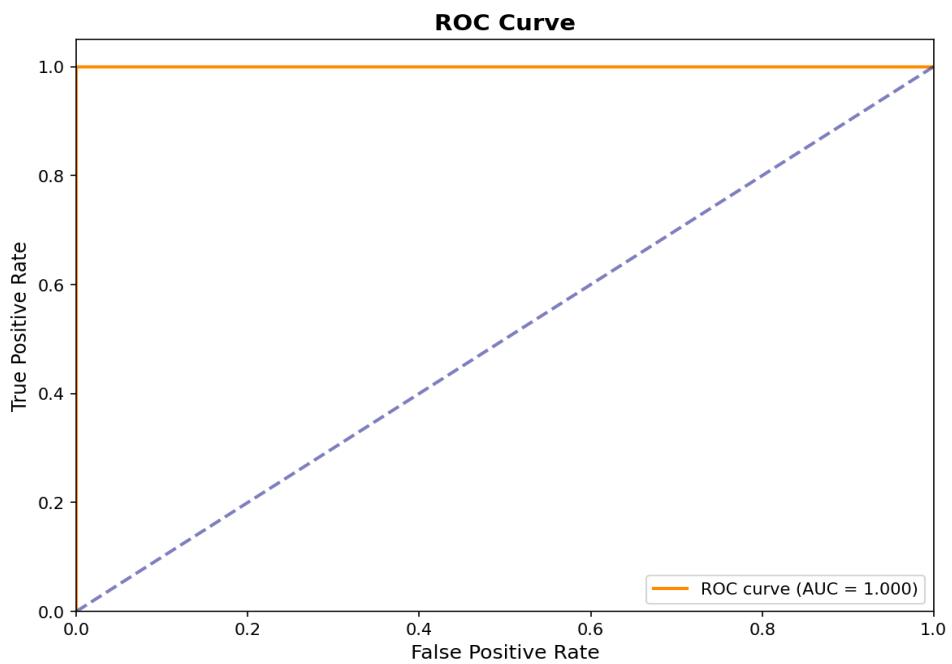
### *Confusion Matrix*



### *Feature Importance*



### *Roc Curve*



## 6. Training Summary

**Total training time:** 6.71 seconds

**Key Events:**

- [SUCCESS] Logistic Regression - Score: 1.0000
- [SUCCESS] Random Forest - Score: 1.0000
- [SUCCESS] Gradient Boosting - Score: 1.0000
- [SUCCESS] Best model: Logistic Regression (Score: 1.0000)

*Report generated by BioMLStudio - AI-Based No-Code Platform for Bioinformatics*  
© 2025 BioMLStudio. All rights reserved.