

Author : S M Tanjimul Haque
Bristol, UK

DNA Promoter Classification Using Machine Learning

Abstract

This project presents a machine learning-based approach to classify DNA sequences as promoter or non-promoter regions. Using the UCI Molecular Biology Promoter Gene Sequences dataset, DNA sequences were preprocessed using one-hot encoding and evaluated across multiple classification algorithms using cross-validation and test set evaluation.

Introduction

Promoters are DNA regions that initiate gene transcription. Identifying promoters is essential in genome annotation, gene regulation studies, and biomedical research. Machine learning provides an efficient computational alternative to traditional laboratory-based promoter detection.

Dataset Description

The dataset contains 106 DNA sequences of length 57 nucleotides. Each instance is labeled as promoter (+) or non-promoter (-).

Methodology

DNA sequences were converted into structured format by splitting nucleotides into positional features. One-hot encoding transformed categorical nucleotides (A, T, G, C) into numeric features. The dataset was split into 75% training and 25% testing sets using stratified sampling.

Models Used

The following algorithms were implemented: KNN, Gaussian Process, Decision Tree, Random Forest, Neural Network (MLP), AdaBoost, Naive Bayes, and Support Vector Machines (Linear, RBF, Sigmoid).

Evaluation

Models were evaluated using 10-fold cross-validation and test set metrics including Accuracy, Precision, Recall, and F1-score.

Conclusion

Kernel-based methods and ensemble models demonstrated strong performance in high-dimensional feature space. The project shows that machine learning can effectively assist promoter prediction tasks.

References

- [1]UCI Machine Learning Repository – Molecular Biology (Promoter Gene Sequences) Dataset
- [2]Scikit-learn Documentation
- [3] Bishop, C. M. (2006). Pattern Recognition and Machine Learning
- [4] Alpaydin, E. (2020). Introduction to Machine Learning