Identification of DNA Exon-Intron Boundaries

Team Name- Huh?

Prachi Bindal- 220102071 Krish Mangal- 220102051

Kushagra Singh Sisodia- 22102052 Tanu Siwach- 220108059

Problem Statement

The challenge presented in this dataset involves identifying specific points on a DNA sequence known as splice junctions. These junctions are where unnecessary sections of DNA are removed during the protein creation process in more complex organisms. The task at hand is to accurately detect, based on a given DNA sequence, the transitions between exons (segments retained after splicing) and introns (segments spliced out). This task comprises two main objectives: identifying boundaries between exons and introns (referred to as EI sites), and identifying boundaries between introns and exons (referred to as IE sites).

Methodology

Input: The input has 181 columns representing 60 nucleotides and a last column representing the class.

The class itself has 3 types :

1 represents: EI: Exon-Intron boundaries (donors)

2 represents: IE: Intron-Exon boundaries(recipients)

3 represents: Neither: Neither exon nor intron

Split data set: We have split the data into training and testing data in other words, kept 80% as training data and 20% as testing data

Methodology

We then utilize **Stacking** implemented on two base models that are **Random Forest** and **Gradient Boosting**.

The stacking methodology will use **Logistic Regression** further to combine the predictions from both base models and fit them to the training data output.

Output

After this predictions made on the testing data provide the classifications of the nucleotide chain.

Predictions for the testing data:

	Actual_Class	Predicted_Class
1029	0	0
1001	0	0
785	2	2
411	2	2
1105	1	1
...	...	...
2623	2	2
693	0	0
2465	2	2
3022	2	2
1356	2	2

Accuracy with Stacking: 0.9655172413793104

Classification Report with Stacking:

	precision	recall	f1-score	support
0	0.94	0.98	0.96	153
1	0.94	0.94	0.94	168
2	0.99	0.97	0.98	317
accuracy			0.97	638
macro avg	0.96	0.96	0.96	638
weighted avg	0.97	0.97	0.97	638

Thank You