Capstone Project Tasks

Objective:

To implement a bioinformatics-based data encoding and feature extraction pipeline, transforming numeric data into biological representations and computing statistical features from nucleotide and amino acid sequences.

Task 1: Numeric Data to Binary Conversion

Convert the given numeric dataset into its equivalent binary representation. Each numeric value will be transformed according to standard binary encoding rules.

Task 2: Assignment of Values to Nucleotides (A, G, C, T) Using WCC Rule

Assign binary values to the four DNA nucleotides — Adenine (A), Guanine (G), Cytosine (C), and Thymine (T) — based on the Watson-Crick Complementary (WCC) rule.

- Example: A ↔ T and G ↔ C, with complementary binary values assigned accordingly.

Task 3: Binary-to-Nucleotide Sequence Conversion

Map the binary data obtained in Task 1 to nucleotide bases (A, G, C, T) using the assigned binary values defined in Task 2, thereby forming a DNA sequence.

Task 4: DNA to RNA Conversion

Perform transcription by converting all occurrences of Thymine (T) to Uracil (U) in the nucleotide sequence, resulting in an RNA sequence.

Task 5: Formation of Codons

Group the RNA sequence into triplets of nucleotides (groups of three bases). Each triplet, known as a codon, represents a potential amino acid.

Task 6: Codon Combination Probability (Feature 1)

Determine the probability of occurrence for each of the 64 possible codon combinations within the RNA sequence. This computed distribution represents Feature 1.

Task 7: Codon-to-Amino Acid Conversion

Translate each codon into its corresponding amino acid using the standard genetic code. Since multiple codons can map to the same amino acid, group molecularly similar codons together to represent 23 unique amino acids.

Task 8: Amino Acid Occurrence Probability (Feature 2)

Compute the probability of occurrence for each of the 23 amino acids obtained in Task 7. This probability distribution forms Feature 2.

Task 9: Feature Combination

Combine the two derived features — Feature 1 (Codon occurrence probabilities) and Feature 2 (Amino acid occurrence probabilities) — to obtain the final composite feature:

- Feature 3 = Feature 1 + Feature 2