Final Project First Attempt

CMSC 435 - Introduction to Data Science

Group 9

Members:
Noah Weingand
Trisha Quiroga Sanchez
Chloe Adzima
John Kelley

Date: October 27, 2020

Section 1: Knowledge Discovery Process

- Researched protein sequences and amino acids
- Developed a Python script using the Biopython package to read in the data to calculate length and molecular weight as features and generate a csv file with those features for each data object.
- Features
    - Length
        - Number of characters (AKA amino acids) in sequence.
    - Molecular Weight
        - Wrote a Python script to calculate using BioPython. If the sequence contained an X or Z, I predicted it as missing value since BioPython would throw an error for sequences with those amino acids. Otherwise, the sequence's molecular weight was calculated.
    - Most Frequent Amino Acid
        - Wrote Python script to removed non-natural residues from nonDRNA objects such as X, B, Z, U, since Pfeature cannot handle these values
        - Used Pfeature to find Amino Acid compositions for each data object
        - Wrote Python script to find the Amino Acid with the highest composition for each protein sequence
- Classification Algorithm
    - Chosen
        - Decision Tree
            - With default parameters
    - Attempted
        - Linear SVM

<u>Section 2:</u> Model Results

Accuracy: 89.08%

predicted

actual	DNA	RNA	DRNA	nonDRNA
DNA	8	4	0	379
RNA	7	6	0	510
DRNA	0	0	0	22
nonDRNA	17	21	0	7821

DNA MCC:

- TP (True DNA): **8**
- FP (False DNA): 7+17= **24**
- TN: 7821 + 6 + 510 + 22 + 21 = **8380**
- FN: 4 + 379 = **383**
- MCC: (8 * 8380) - (24 * 383) / √(8+24)(8+383)(8380+24)(8380+383) = **0.060**

RNA MCC:

- TP: **6**
- TN: 8 + 379 + 17 + 22 + 7821= **8247**
- FP: 7 + 0 + 510 = **517**
- FN: 4 + 0 + 21 = **25**

- MCC = **0.034**

DRNA MCC:

- TP: **0**
- TN: 8 + 4 + 7 + 6 + 17 + 21 + 379 + 510 + 7821 = **8773**
- FP: **0**
- FN: **22**
- MCC = (0 * 8773) - (0 * 22) / √(0 * 0)(0 * 22)(8773 * 0)(8773 * 22) = **0**

nonDRNA MCC:

- TP: **7821**
- TN: 8 + 4 + 7 + 6 = **25**
- FN: 17 + 21 = **38**
- FP: 379 + 510 + 22 = **911**
- MCC = **0.07999**