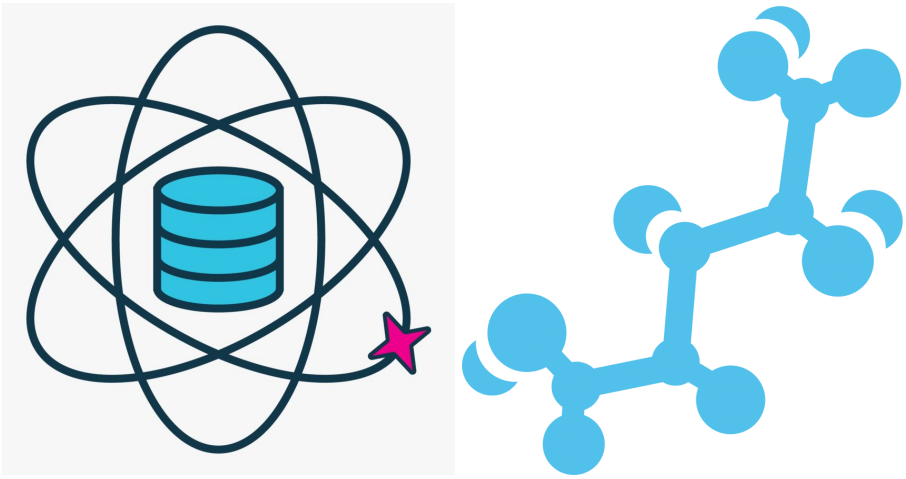
CMSC 435
Intro to Data Science

Group 14 Project



Basima Zafar
Daniel West
Josh Lopez
Hunter Arndorfer

October 27, 2020

Preprocessing the data:
As a group we chose to use Pfeature to generate our feature sets. In order to upload our training data to this resource, we needed to convert the .txt file to a .csv and also remove the Class identifier from each row in the training data. We wrote a simple python script that would do the following:

- Convert .txt file to .csv using Pandas
- Strip classifiers off csv and save as new file
- Run new csv through Pfeature
- Pfeature returned a dataset of all sequences and their features in csv format
- Append classifiers back onto new dataset

Our first attempt:

- Our first attempt consisted of gathering sequence features from the website Pfeature and then running that dataset through our model in RapidMiner.

- We generated a total of 21 features for our dataset.
- These features include:
    - AAC_A - Amino acid composition of Alanine
    - AAC_C - Amino acid composition of Cysteine
    - AAC_D - Amino acid composition of Aspartic acid
    - AAC_E - Amino acid composition of Glutamic acid
    - AAC_F - Amino acid composition of Phenylalanine
    - AAC_G - Amino acid composition of Glycine
    - AAC_H - Amino acid composition of Histidine
    - AAC_I - Amino acid composition of Isoleucine
    - AAC_K - Amino acid composition of Lysine
    - AAC_L - Amino acid composition of Leucine
    - AAC_M - Amino acid composition of Methionine
    - AAC_N - Amino acid composition of Asparagine
    - AAC_P - Amino acid composition of Proline
    - AAC_Q - Amino acid composition of Glutamine
    - AAC_R - Amino acid composition of Arginine
    - AAC_S - Amino acid composition of Serine
    - AAC_T - Amino acid composition of Threonine
    - AAC_V - Amino acid composition of Valine

- AAC_W - Amino acid composition of Tryptophan
- AAC_Y - Amino acid composition of Tyrosine
- Class - Object classifier (DNA, RNA, DRNA, nonDRNA)

- In RapidMiner, we chose the Decision Tree as our classification algorithm with 5-fold cross validation.

$$MCC = \frac{TP*TN - FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

Confusion Matrix

	DNA	RNA	DRNA	nonDRNA
DNA	5	3	0	6
RNA	9	98	0	31
DRNA	0	0	0	0
nonDRNA	377	422	22	7822

Outcomes

	TP	FN	TN	FP
DNA	5	386	8395	9
RNA	98	425	8232	40
DRNA	0	22	8773	0
nonDRNA	7822	37	115	821

Sensitivity

DNA	1.278772379
RNA	1.278772379
DRNA	0
nonDRNA	99.52920219

Specificity

DNA	99.89290814
RNA	99.51644101
DRNA	100
nonDRNA	12.28632479

Predictive

DNA	95.50881182
RNA	94.71290506
DRNA	99.74985787
nonDRNA	90.24445708

MCC

DNA	0.06057655939
RNA	0.3473782128
DRNA	0
nonDRNA	0.2795893435

Average MCC: 0.1718860289

Accuracy: 90.10801592