Abhinav **Suresh**
Mohnish **Agrawal**
Rhythm **Patel**

IQB Assignment 2

README

Along with predicting, this code also finds the Train & Test accuracy with K-Fold method and 5 folds.

Syntax Format:

```
python3 final.py -o <output File> -type {aminoacid | dipepeptide | tripeptide} -model {SVC | RFC} { {-gamma | -n_estimators} <positive value> | -optimalParameters }
```

Understanding of Syntax:

- Type can only be ‘aminoacid’, 'dipeptide' or ‘tripeptide’. These are the different ways of parsing the input peptide.
- Models:
  - SVC: Support Vector Classifier
  - RFC: Random Forest Classifier
- -gamma and -n_estimators are the parameters we pass into these models.
- -optimalParameters is a mode where it runs the model multiple times and plots the accuracy for the parameters. This way, we can find out the optimal gamma/n_estimators for a particular model.

Please Note:

- gamma can only be used if model is SVC.
- n_estimators can only be used if model is RFC.
- gamma & -n_estimators have to be positive values
- outputFile is preferred with .csv format
- In -optimalParameters, 3 inputs will be asked: start, end, step indexes for the parameters.
- If -optimalParameters mode is used, -gamma <val> cannot be used for obvious reasons.

Best Models (Acc to avg test accuracy):

1. Model: RFC
   Type: Tripeptide
   n_estimators: 700
   CSV stored in: best1.csv
   Code in: best1.py

```
python3 best1.py OR
python3 final.py -o best1.csv -type tripeptide -model RFC
-n_estimators 700
```

```
Output exported in best1.csv
Avg Train Accuracy:  1.0
Avg Test Accuracy:  0.8901960784313726
```

2. Model: SVC
   Type: Tripeptide
   Gamma: 15
   CSV stored in: best2.csv
   Code in: best2.py

```
python3 best2.py OR
python3 final.py -o best2.csv -type tripeptide -model SVC -gamma
15
```

```
Output exported in best2.csv
Avg Train Accuracy:  0.9859803921568627
Avg Test Accuracy:  0.8741176470588237
```

3. Model: SVC
   Type: Dipeptide
   Gamma: 91
   CSV stored in: best2.csv
   Code in: best2.py

```
python3 best3.py OR
python3 final.py -o best3.csv -type dipeptide -model SVC -gamma
91
```

```
Output exported in best3.csv
Avg Train Accuracy:  0.9981372549019607
Avg Test Accuracy:  0.8631372549019607
```

Example Syntax of optimalParameters:

```
python3 final.py -o a.csv -type aminoacid -model SVC
-optimalParameters

      Enter start parameter: 40
      Enter end parameter: 200
      Enter step parameter: 30
      Gamma : 40
      Avg Train Accuracy:  0.9235294117647058
      Avg Test Accuracy:  0.851764705882353
      Gamma : 70
      Avg Train Accuracy:  0.9575490196078432
      Avg Test Accuracy:  0.8623529411764705
      Gamma : 100
      Avg Train Accuracy:  0.9748039215686275
      Avg Test Accuracy:  0.8619607843137255
      Gamma : 130
      Avg Train Accuracy:  0.9858823529411765
      Avg Test Accuracy:  0.8564705882352941
      Gamma : 160
      Avg Train Accuracy:  0.9920588235294119
      Avg Test Accuracy:  0.8564705882352941
      Gamma : 190
      Avg Train Accuracy:  0.993921568627451
      Avg Test Accuracy:  0.8525490196078431
```