Big Data Mining in Healthcare (BDMH)
Assignment – 2

Submitted by:

Riya garg -MT22058
Pranshu Patel-MT22117
Anand Singh Rathore-PhD22201

How to run the file ?

Command line input: To run the given python files, the input command at the command line terminal is as:
```
python Group_9.py train.csv test.csv
```

Output generated : run_15.csv

Highest Accuracy obtained (public score):0.79569

Approach:

The intermediate steps used for developing a model to classify the variable length of peptides are as discussed below: -

1.**Feature Pre-Processing and Generation**: - First we checked if data is clean and then we apply different preprocessing on the train and test data. Then for feature generation we applied 2 methods:

METHOD 1:

The features are generated by dividing the frequency of one character in a sequence by the length of the sequence.Then we also calculated features by dividing the frequency of two contiguous amino acids in a sequence by the length of the sequence.

METHOD 2:

We used features and calculated AAC and DPC of the given sequences in the train and test dataset and merged them into a new dataset using pfeature_comp.py

2. Then we split the data into train and test data

3.**Model Training:** We applied various models like XGB , random forest ,SVM ,MLP ,ExtraTreeClassifier ,on both the features generated by method1 and method2.
We got following accuracy on kaggle of each model using given feature selection method:

Method1:

ExtraTreeClassifier-0.79569
Random forest :0.77538

MLP:0.68207
SVM: 0.7644

Method2:

ExtraTreeClassifier: 0.77568
Random forest : 0.77692

4. So the Best model(extra tree classifier on method 1) was trained and the final predictions are proposed using predict_proba() on the test file supplied.

Requirements:

The requirements for the virtual environment are as listed below: -

- For Python Version = 3.10.7
- joblib==1.2.0
- numpy==1.24.2
- pandas==1.5.3
- python-dateutil==2.8.2
- pytz==2022.7.1
- scikit-learn==1.2.2
- scipy==1.10.1