Assignment - 1
Prediction of Interacting Patterns

Group: 58
Best Score: 0.64793
Anuneet Anand (2018022)
Pankil Kalra (2018061)
Akanksha Arora (PhD20208)

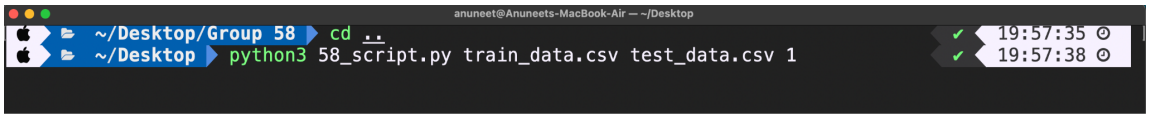
Files

- 58_script.py: Python code file with all the models.
- train_data.csv: Provided training dataset.
- test_data.csv: Provided test dataset.
- Submission_x.csv : 3 Files corresponding to 3 models with probabilities.
- Predictions_x.csv : 3 Files corresponding to 3 models with binary values.

Usage

- Requirements: imblearn, scikit_learn, pandas
- The given script takes the training dataset, test dataset and model number of predefined models as arguments and accordingly generates the predictions as CSV files.

```
Usage: python3 58_script.py <train_data> <test_data> <model>
```



- The train_data and test_data files should be present in the same directory.
- We generate Submission_x.csv files for predicted probabilities (used on Kaggle) and also Prediction_x.csv files for predicted 0/1 values.
- The generated output files are stored in the same directory.
- Valid values for models are 1, 2, 3 (corresponding to our 3 best models) and 0 to run all the 3 models.

Models

- **Model 1:** We used One Hot Encoding for encoding the protein sequences which gives a matrix with 357 features (17*21). This helps us to capture the position-specific presence of different amino acids in the sequence. This was then subjected to undersampling as we have imbalanced classes in the data. We used a Stacking Classifier with Random Forest Classifier, MLP Classifier and KNN as base estimators and LogisticRegression as the final estimator. The predicted probabilities are stored in Submission_1.csv which corresponds to 58_Submission_3.csv on Kaggle.
  Public Score : 0.64764
  Private Score : 0.64793

- **Model 2:** We used First order Dipeptide Composition for encoding the protein sequences which gives a matrix with 400 features (20*20). This helps us to capture information on % composition of various Dipeptides in our sequence. We used a Balanced Bagging Classifier with SVC (rbf kernel) as base estimator. Balanced Bagging Classifier handled the class imbalance and helped in bagging while SVC served as a good base predictor. The predicted probabilities are stored in Submission_2.csv which corresponds to DPC_SVM_BBC_100_prob.csv on Kaggle.
  Public Score : 0.60984
  Private Score : 0.63086

- **Model 3:** We used First order Dipeptide Composition for encoding the protein sequences which gives a matrix with 400 features (20*20). This helps us to capture information on % composition of various Dipeptides in our sequence. We used a Balanced Bagging Classifier with Random Forest Classifier as base estimator. Balanced Bagging Classifier handled the class imbalance and helped in bagging while Random Forest served as a good base predictor. The predicted probabilities are stored in Submission_3.csv which corresponds to Submission_1.csv on Kaggle.
  Public Score : 0.60400
  Private Score : 0.61567