BDMH Assignment 2 : Classification of Proteins

Group_4
Rohan Kulkarni (2020537)
Deep Sharma(2020370)
Cyrus Monteiro (2020368)

Overview

This script trains the model on the given data and outputs the test set’s results and prints the evaluation metrics

Prerequisites

Before running the script, ensure you have installed all the dependencies mentioned in the 'requirements.txt' file.
You can use pip install -r requirements.txt or pip install “package name”

Note

Pfeatures - features were extracted using the web portal

Command Line Options

```
python predict.py -tr train.csv -te test.csv -o output.csv
```

When running the code on command line or local machine, make sure the python version installed in your machine is Python 3.10 or above.

We used:

AG (AutoGluon)

This model trains multiple machine learning models using AutoGluon and makes predictions on the test data.

-
Inputs –
Training Data: train.csv
Testing Data: test.csv
Test File: test.csv
-
Outputs –
Output Folder: autog folder

Functionality: The script reads three CSV files. The features include TF-IDF and One Hot Encoding and physico-chemical properties, features extracted using BioPython. The Model then calls the TabularDataset class of AutoGluon to load the features, divides the training data into train and validation set and then calls the TabularPredictor class to make predictions.

- Finally outputs the predictions in a CSV File.

Oversampling was done using Borderline SMOTE