MLBA Assignment 2 - Group 71

README

● Input Files:

The program takes paths to the training data as the user input. We have used the files provided on Kaggle as our training and testing data. For training data: train.csv For testing data: test.csv.

● Output Files:

The program outputs a CSV file, predictions.csv containing the predicted binary class labels (0,1) for the training data. The name of the prediction file submitted on Kaggle was different, though. Prediction file submitted on Kaggle: RFC_New.csv

● Steps To Run the Python File:

Make sure that all dependencies and libraries used by this program are installed on your system.

A list of the required libraries are:

➔ Numpy

➔ Pandas

➔ Sklearn

➔ TensorFlow

➔ Seaborn

➔Matplotlib

Use the following steps to run the code:

Dependencies:

1. Sklearn
2. Tensorflow
3. Numpy
4. Pandas

Command Line Options:

1. Type the command “python3 group71_code.py <train_csv_path> <test_csv_path>” in the same directory as the notebook.
2. Change the file paths to your file paths.
3. Install all dependencies mentioned above.
4. After this, we will run all the cells and get a CSV file as the output “predictions.csv.”

Why Run?

To predict the labels for the protein sequences.

Model Info:

The Random Forest Classifier, with feature selection (SelectKBest), retains 28 features and uses 100 decision trees.

Further models tried:

1. Convolutional Neural Network
2. Random Forest
3. Multi-Layer Perceptron
4. XGBoost
5. Bagging

Some of the results from the other models on default parameters, just to check the best performing model.

	Accuracy
k-Nearest Neighbors	0.605263
Support Vector Machine	0.578947
Linear SVM	0.644737
RBF SVM	0.578947
Gaussian Process	0.578947
Decision Tree	0.539474
Extra Trees	0.565789
Random Forest	0.552632
Extra Forest	0.605263
AdaBoost	0.565789
Gaussian Naive Bayes	0.486842
LDA	0.486842
QDA	0.552632
Logistic Regression	0.605263
SGD Classifier	0.592105
Multilayer Perceptron	0.434211
Voting Classifier	0.644737

The best accuracy given was the **Random Forest Classifier** upon further hyper-parameterization. The accuracy given by the classifier was 85.1% upon Kaggle submission.