-Priyanka Boral (MT19127)-Reecha Kumari Giri(MT19134)- Akanksha Dewangan(MT19049)

MLBA ASSIGNMENT-1

(Group-9)

- ***Readme***:

  A. To run the code named Priyanka_Boral_MT19127_SVM(73_64).ipynb file with best accuracy(leaderboard score= 0.73664)

(i) First merged data named "merged_train.csv and merged_test.csv" are to be uploaded.

(ii) Run each and every cell of the kernel. (Documentation is done in the notebook itself).

(iii) On running the last cell, output file with .csv extension could be downloaded which is the class prediction of the test data (prediction of proteins in DNA-binding and non-binding proteins.

Output file submitted using this code: svm(73_64)_output.csv

B. To run the code named Priyanka_Boral_MT19127_votingclassifier(71_028).ipynb file with accuracy(leaderboard score= 0.71028)

(i) First merged data named "merged_train.csv and merged_test.csv", "dipeptide_matrix and dipeptide_test_matrix" are to be uploaded.

(ii) Run each and every cell of the kernel. (Documentation is done in the notebook itself).

(iii) On running the last cell, output file with .csv extension could be downloaded which is the class prediction of the test data (prediction of proteins in DNA-binding and non-binding proteins.

Output file submitted using this code: votingClassifier_(71_028).csv

C. To run the code named Priyanka_Boral_MT19127_SVM_dipep(71_4).ipynb file with accuracy(leaderboard score= 0.714)

(i) First merged data named "dipeptide_matrix and dipeptide_test_matrix" are to be uploaded.

(ii) Run each and every cell of the kernel. (Documentation is done in the notebook itself).

(iii) On running the last cell, output file with .csv extension could be downloaded which is the class prediction of the test data (prediction of proteins in DNA-binding and non-binding proteins.

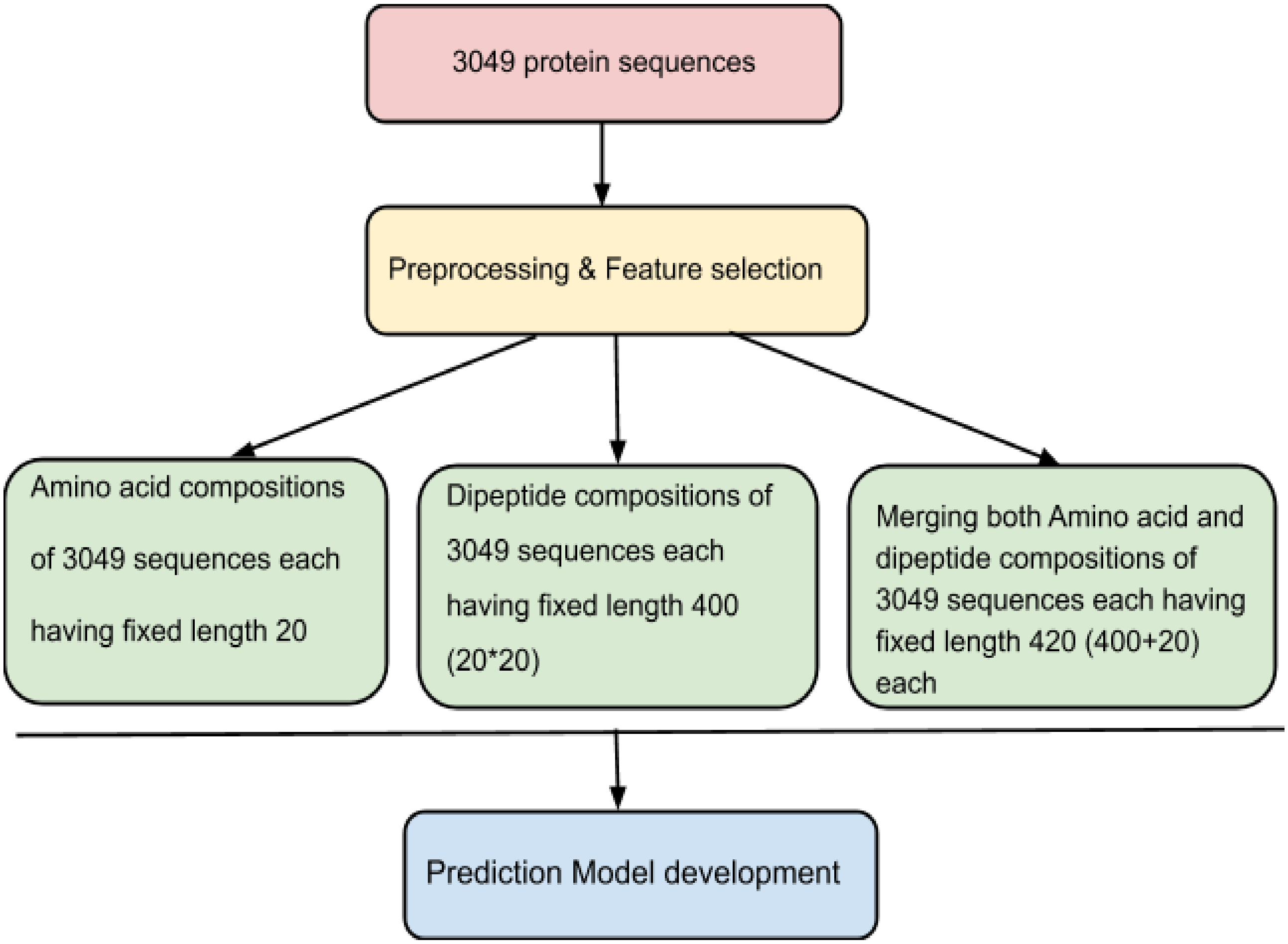
Output file submitted using this code: svm_dipep(71_4)_output.csv

Report contains information regarding all the techniques used for this assignment and analysis.

★ **Aim**: To develop machine learning models to classify proteins in DNA-binding and non-binding proteins.

★ Preprocessing and Methodology:

Complete description through flow chart


3049 protein sequences
Preprocessing & Feature selection
Amino acid compositions of 3049 sequences each having fixed length 20
Dipeptide compositions of 3049 sequences each having fixed length 400 (20*20)
Merging both Amino acid and dipeptide compositions of 3049 sequences each having fixed length 420 (400+20) each
Prediction Model development

Preprocessing and Feature Selection:

Amino acid and dipeptide sequences are generated from p-feature web server.

1. After converting train and test data to compositions of fixed length, since it is in -1 & 1 form. So, labels are transformed in the form of 0 and 1, by converting -1 to 0.

(i) Amino acid Composition formula:

AACi= Ri/L

Where, AACi is amino acid composition of residue type I; Ri and L are number of residues of type I and length of sequence.

(ii) **Dipeptide Composition Formula:**

DPCi j = Di j/ L − j

Where DPCj i is the fraction or composition of dipeptide of type i for jth order. D j i and L are the number of dipeptides of type i and length of a protein. Here higher order dipeptide D j i is made of residue Ri and Ri+j where value of j is 2 or more. In case j is equal to 1 then dipeptide is called traditional dipeptide.

Feature Selection:

(i) SelectKBest RFE(Recursive feature elimination) for 215 feature extraction

(ii) SelectKBest,f_classif sel_f = SelectKBest(score_func=f_classif, k=215)

2. Various models are developed for prediction of proteins in DNA-binding and non-binding proteins.
3. Finally after obtaining predictions from the developed model, they are converted in -1 and 1 form i.e., 0 is converted to -1.

Models developed using deep learning techniques:

1. Support Vector Machine(SVM)

2. Random Forest

3. **Extra Tree Classifier**

4. **Adaboost**

5. **Naive Bayes**

7. **XGBoost**

8. **KNN**

9. **Ensemble Classifiers (Voting Classifier(Extra tree classifier Random forest + xgboost), (adaboost + SVM + Extra tree classifier + random forest), (SVM+ Extratreeclassifier + Random Forest)**

All the above models are giving different results on using various parameters and on various runs. Best validation scores which are obtained on the parameters are mentioned in the below table.

Classifiers	**Validation Accuracy**	**Parameters**	**Composition**	**Leaderboard**
1. **SVM**	70.8	`(kernel='linear', C=1, gamma=3).fit(X_train, y_train)`	Amino Acid	0.68224
2. **SVM** **(selected for final submission)**	70.2	`(kernel='linear', C=3, gamma=3).fit(X_train, y_train)`	Dipeptide	0.714

3. **SVM**	71.67	(kernel='linear', C=3, gamma=3).fit(X_train, y_train)	Amino-acid + Dipeptide	0.7121
4. **SVM with feature selection(selectKBest= fannova)**	72.8	(kernel='rbf', C=3, gamma=’scale’).fit (X_train, y_train)	Amino acid+Dipeptide	0.7364
5. **Random Forest** **(selected for final submission)**	71.7	(n_estimators=135, random_state=None, n_jobs=4, criterion='entrop)	Amino acid	0.59-0.71214
6. **Extra Tree Classifier**	71.13	(n_estimators=num_trees, criterion='gini',min_samples_split=2, max_features=max_features)	Amino acid	0.65-0.712
7. **Adaboost**	67.07	(n_estimators=num_trees, random_state=seed)	Amino acid	***

8. Naive Bayes	66.7	GaussianNB()	Dipeptide	***
9. KNN	55.23	KNeighborsClassifier(n_neighbors=21)	Amino acid	***
10. XGboost	68.13	n_estimators=110, max_depth=70,max_features=110	Amino acid	***
11. Major Voting Classifier() **(selected for final submission)**	71.7	('gnb', clf),('ext', model),('rnd',rdf),('ada',model2)], voting='soft')	Amino acid+Dipeptide	0.71028

- For validation accuracy, cross validation K-fold (10-fold) has been used.

- In the above table besult results of each models are shown but is checked through different parameters like by changing n_estimators, max_features, max_depth, min_samples_split, C=1 to 5 to 10, gamma='entropy', 'gini', 'sigmoid', 'linear', class-weight='balanced' etc.

[Note: *** denotes that the model which we used were not submitted in kaggle because they are not giving a good cross-validation score.]

CONCLUSION:

- From above observation, it is concluded that in our case, **SVM with feature selection of 215 features(selectKBest= f-anova) on** (kernel= 'rbf', C=3, gamma='scale', class weight= 'balance').fit(X_train, y_train) parameter on the merged data of dipeptide and amino acid is giving best result with cross **validation accuracy= 72.8%** and **leaderboard score=0.7364**.

- 2nd best result is obtained from the ensemble method of **validation accuracy= 71.7%** and **leaderboard score= 0.71028** using **voting classifier on (SVM on amino acid, Extra tree classifier, SVM on merged data , dipeptide data, feature selection, SVM on merged data)** on merged data of dipeptide,amino acid and independent dipeptide data with feature selection of 215 features (selectKbest=f-anova).

- In our case, **SVM on** (kernel= 'rbf', C=3, gamma='scale', class weight= 'balance').fit(X_train, y_train) parameter on the dipeptide data is giving the best result with cross **validation accuracy= 70.2%** and **leaderboard score=0.714**.

Note: The results may vary on different parameters and on various runs.

-Priyanka Boral (MT19127)-Reecha Kumari Giri(MT19134)- Akanksha Dewangan(MT19049)