HUST

TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI

HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

ONE LOVE. ONE FUTURE.

TRƯỜNG ĐẠI HỌC
BÁCH KHOA HÀ NỘI
HANOI UNIVERSITY
OF SCIENCE AND TECHNOLOGY

Tin sinh học

Xây dựng mô hình dự đoán peptides chống tạo mạch

Giảng viên hướng dẫn: TS. Nguyễn Hồng Quang

Tống Mạnh Đạt, MSSV : 20173008
Lớp : KTMT 06, K62
Đặng Quang Anh : 20172942
Lớp : KTMT07-K62

ONE LOVE. ONE FUTURE.

1. Giới thiệu

2.Mô tả bài toán

3.Đề xuất mô hình

4.Thực hiện hệ thống

5.Các thử nghiệm

1.Giới thiệu

- Dù công nghệ phát triển rất nhanh, nhưng Ung thư vẫn là một trong các bệnh nan y khó chữa và gây tử vong nhất nếu không được phát hiện ra sớm.
- Peptit được coi là một liệu pháp điều trị quan trọng đang được thử nghiệm đối với các bệnh phụ thuộc vào quá trình tạo mạch bởi độc tính thấp cùng hiệu quả cao. Các peptides chống tạo mạch nhiều triển vọng trong các nghiên cứu tiền lâm sàng và lâm sàng đối với bệnh ung thư . Do đó, việc dự đoán được peptide chống tạo mạch là những ứng cử viên đầy hứa hẹn trong điều trị ung thư.

1.Giới thiệu

- Các nghiên cứu đã có :

Phương pháp	Mô hình phân lớp	Đặc trưng của Sequence	Independent Test	Web Server
AntiAngioPred	SVM	AAC (20)	Có	Có
Blanco et al.'s method	glmnet	AAC, DPC, TC (200)	Không	Không
AntAngioCOOL	PART	PseAAC, k-mer composition, RAAC, PCP, AC (2,343)	Không	Không
TargetAntiAngio	RF	AAC, PseAAC, Am-PseAAC (48)	Có	Có
Nghiên cứu này	RF, Ada boost, extratrees	AAC,dpc	Không	Không

2.Mô tả bài toán

- 2.1.Chi tiết bài toán :

Đầu vào : cho mỗi chuỗi peptides ngắn. ví dụ : ADNWQSFDRWKDH.

Định dạng dữ liệu:

>AA135

YTMNPRKLFDY

>neg1

ADNWQSFDRWKDH

Với “AA135”, “neg1” là tên các peptide và “YTMNPRKLFDY”, “ADNWQSFDRWKDH” là trình tự của peptide tương ứng.

Đầu ra : dự đoán có phải có chức năng chống tạo mạch (anti-angiogenic peptides: Antiangio - 1) hay không (non-antiangiogenic peptides: Negative - 0).

2.Mô tả bài toán

- 2.1.Tập dữ liệu :

Tập dữ liệu gồm 2 file fasta được lấy từ nghiên cứu TargetAntiAngio - Dự đoán và phân tích peptides chống tạo mạch.

benchmarkdataset.fasta làm tập Train : 135 peptide sequences Thuộc lớp peptide chống tạo mạch và 135 peptides ngẫu nhiên sử dụng làm peptides không chống tạo mạch.

NT15dataset.fasta làm tập Test: it chứa 99 peptides chống tạo mạch và 101 Không chống tạo mạch.

Dataset	Benchmarkdataset - tập train		NT15dataset - tập test	
	Anti-anigo	Non-anti-anigo	Anti-anigo	Non-anti-anigo
Tổng dữ liệu	135	135	99	101

2.Mô tả bài toán

- 2.2.Tập dữ liệu :

Cách chia tập dữ liệu khác : Gộp 2 tập benchmarkdataset và NT15 sau đó chia ngẫu nhiên thành 2 tập dữ liệu train và test với tỷ lệ (8 -2)

Dataset	Tập train		Tập test	
	Anti-anigo	Non-anti-anigo	Anti-anigo	Non-anti-anigo
Tổng dữ liệu	187	189	47	47

2.Mô tả bài toán

- 2.2.Các độ đo :

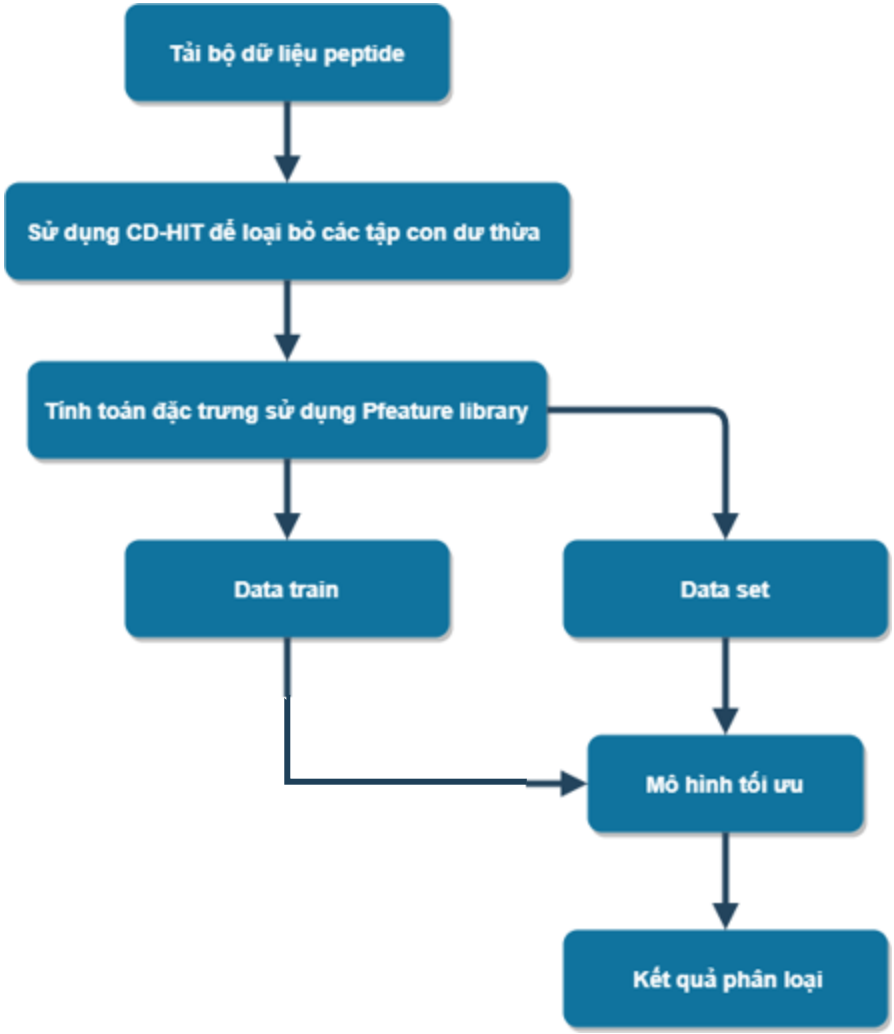
$$\mathrm{Ac} = \frac{\mathrm{TP} + \mathrm{TN}}{(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN})}$$

$$\mathrm{Sn} = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FN})}$$

$$\mathrm{Sp} = \frac{\mathrm{TN}}{(\mathrm{TN} + \mathrm{FP})}$$

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})\,(\mathrm{TP} + \mathrm{FN})\,(\mathrm{TN} + \mathrm{FP})\,(\mathrm{TN} + \mathrm{FN})}}$$

3.Đề xuất mô hình



- ✓ Tải bộ dữ liệu peptide bao gồm bộ Antiangio - peptide chống tạo mạch và negative - không phải peptide chống tạo mạch
- ✓ Sử dụng CD-HIT để loại bỏ các tập con dư thừa
- ✓ Tính toán các đặc trưng bằng thư viện Pfeature:
  - Xác định các hàm để tính toán các đặc trưng khác nhau
  - Tính toán đặc trưng cho cả tập Antiangio và tập negative, kết hợp lại và thêm 1 cột đánh nhận.
- ✓ Áp dụng mô hình tối ưu đưa ra kết quả phân loại

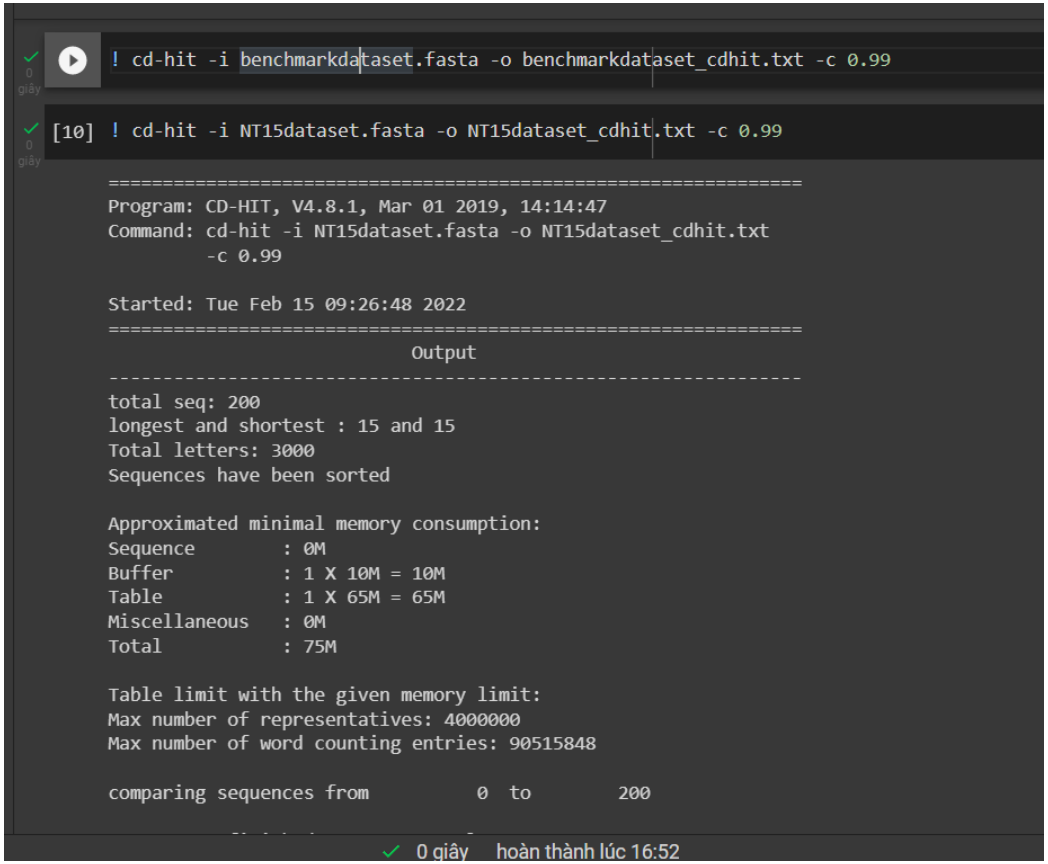
4.Thực hiện hệ thống:

Tải về các tập dữ liệu :

```
! wget https://raw.githubusercontent.com/tmd22121999/Antimicrobial_Peptide/main/NT15dataset.fasta
! wget https://raw.githubusercontent.com/tmd22121999/Antimicrobial_Peptide/main/Class%20angio%20and%20negative%20NTCT.csv
! wget https://raw.githubusercontent.com/tmd22121999/Antimicrobial_Peptide/main/Class%20angio%20and%20negative.csv
! wget https://raw.githubusercontent.com/tmd22121999/Antimicrobial_Peptide/main/benchmarkdataset.fasta
```

```
--2022-02-15 09:26:37--  https://raw.githubusercontent.com/tmd22121999/Antimicrobial_Peptide/main/NT15dataset.fasta
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 185.199.108.133, 185.199.109.133, 185.199.110.133, ...
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|185.199.108.133|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 4944 (4.8K) [text/plain]
Saving to: 'NT15dataset.fasta'

NT15dataset.fasta   100%[===================>]   4.83K  --.-KB/s    in 0s

2022-02-15 09:26:37 (65.9 MB/s) - 'NT15dataset.fasta' saved [4944/4944]

--2022-02-15 09:26:37--  https://raw.githubusercontent.com/tmd22121999/Antimicrobial_Peptide/main/Class%20angio%20and%20negative%20NTCT.
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 185.199.110.133, 185.199.108.133, 185.199.111.133, ...
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|185.199.110.133|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 3322 (3.2K) [text/plain]
Saving to: 'Class angio and negative NTCT.csv'

Class angio and neg 100%[===================>]   3.24K  --.-KB/s    in 0s

2022-02-15 09:26:37 (56.9 MB/s) - 'Class angio and negative NTCT.csv' saved [3322/3322]

--2022-02-15 09:26:37--  https://raw.githubusercontent.com/tmd22121999/Antimicrobial_Peptide/main/Class%20angio%20and%20negative.csv
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 185.199.108.133, 185.199.109.133, 185.199.110.133, ...
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|185.199.108.133|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 2820 (2.8K) [text/plain]
Saving to: 'Class angio and negative.csv'
```

4.Thực hiện hệ thống:



- 4.1.Tiền xử lý dữ liệu

- Ban đầu tập benchmark dataset chứa 257 trình tự peptides thuộc lớp peptides chống tạo mạch. Tập dữ liệu được tiền xử lý bằng cách sử dụng CD-HIT lọc các peptide có trình tự giống nhau trên 99%.

4.Thực hiện hệ thống:

- 4.2.Xử lý đầu vào :
- Đặc trưng AAC:

$$f(a) = {N_a}/{N},\ a \in (A,C,...,W,Y)$$

Với f(a) là đặc trưng AAC của aminoacid loại a trong trình tự

Na = tổng số lượng aminoacid loại a

N = tổng số lượng aminoacid trên toàn bộ trình tự .

- Đặc trưng DPC

$$f(a,b) = {N_{ab}}/{(N-1)},\ \ a,b \in (A,C,...,W,Y)$$

f(a,b) là đặc trưng AAC cúa depeptit loại ab trong trình tự

Nab = tổng số lượng depeptit loại ab

N = tổng số lượng aminoacid trên toàn bộ trình tự .

Xử lí dữ liệu đầu vào

- ✓Tiến hành trích chọn đặc trưng AAC(Amino acid composition) do thư viện Pfeature cung cấp. Đặc trưng này tính toán tỉ lệ xuất hiện của từng amino axit trong 20 amino acid trong chuỗi peptide
- ✓Tính toán đặc trưng AAC cho 2 tập Antiangio và negative, sau đó tổng hợp lại thành 1 file chung và thêm 1 cột gán nhãn

	AAC_A	AAC_C	AAC_D	AAC_E	AAC_F	AAC_G	AAC_H	AAC_I	AAC_K	AAC_L	AAC_M	AAC_N	AAC_P	AAC_Q	AAC_R	AAC_S	AAC_T	AAC_V	AAC_W	AAC_Y	Class
0	15.79	10.53	0.00	5.26	15.79	10.53	5.26	0.00	0.00	5.26	0.00	5.26	5.26	10.53	5.26	0.00	5.26	0.00	0.00	0.00	Antiangio
1	5.88	0.00	5.88	2.94	2.94	2.94	2.94	8.82	17.65	17.65	2.94	5.88	0.00	2.94	5.88	5.88	0.00	5.88	2.94	0.00	Antiangio
2	35.00	0.00	5.00	5.00	0.00	0.00	0.00	0.00	20.00	0.00	0.00	0.00	10.00	5.00	10.00	0.00	5.00	5.00	0.00	0.00	Antiangio
3	15.79	10.53	0.00	0.00	0.00	15.79	0.00	0.00	0.00	0.00	0.00	0.00	0.00	5.26	15.79	21.05	5.26	5.26	5.26	0.00	Antiangio
4	15.79	10.53	0.00	5.26	10.53	10.53	5.26	5.26	0.00	0.00	0.00	5.26	5.26	0.00	5.26	5.26	10.53	0.00	0.00	5.26	Antiangio
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
195	6.67	0.00	0.00	13.33	0.00	6.67	0.00	0.00	6.67	6.67	0.00	0.00	6.67	0.00	26.67	6.67	0.00	13.33	0.00	6.67	Negative
196	0.00	0.00	20.00	13.33	0.00	6.67	0.00	6.67	6.67	20.00	0.00	0.00	0.00	0.00	6.67	6.67	0.00	6.67	0.00	6.67	Negative
197	0.00	0.00	0.00	0.00	6.67	0.00	0.00	6.67	0.00	20.00	6.67	6.67	0.00	6.67	0.00	20.00	6.67	13.33	0.00	6.67	Negative
198	0.00	0.00	0.00	20.00	6.67	20.00	0.00	0.00	0.00	6.67	6.67	0.00	6.67	6.67	6.67	0.00	0.00	6.67	0.00	13.33	Negative
199	6.67	0.00	6.67	0.00	6.67	0.00	0.00	13.33	6.67	26.67	0.00	6.67	0.00	0.00	0.00	13.33	6.67	0.00	0.00	6.67	Negative

468 rows × 21 columns

- ✓ Tiến hành phân tách đặc trưng thành X, cột nhãn thành Y để có thể xây dựng mô hình sau này.

Thực hiện hệ thống

✓Từ bước xử lý dữ liệu đầu vào trên, ta đang có ma trận X và y như sau:

	AAC_A	AAC_C	AAC_D	AAC_E	AAC_F	AAC_G	AAC_H	AAC_I	AAC_K	AAC_L	AAC_M	AAC_N	AAC_P	AAC_Q	AAC_R	AAC_S	AAC_T	AAC_V	AAC_W	AAC_Y	Class
0	15.79	10.53	0.00	5.26	15.79	10.53	5.26	0.00	0.00	5.26	0.00	5.26	5.26	10.53	5.26	0.00	5.26	0.00	0.00	0.00	Antiangio
1	5.88	0.00	5.88	2.94	2.94	2.94	2.94	8.82	17.65	17.65	2.94	5.88	0.00	2.94	5.88	5.88	0.00	5.88	2.94	0.00	Antiangio
2	35.00	0.00	5.00	5.00	0.00	0.00	0.00	0.00	20.00	0.00	0.00	0.00	10.00	5.00	10.00	0.00	5.00	5.00	0.00	0.00	Antiangio
3	15.79	10.53	0.00	0.00	0.00	15.79	0.00	0.00	0.00	0.00	0.00	0.00	0.00	5.26	15.79	21.05	5.26	5.26	5.26	0.00	Antiangio
4	15.79	10.53	0.00	5.26	10.53	10.53	5.26	5.26	0.00	0.00	0.00	5.26	5.26	0.00	5.26	5.26	10.53	0.00	0.00	5.26	Antiangio
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
195	6.67	0.00	0.00	13.33	0.00	6.67	0.00	0.00	6.67	6.67	0.00	0.00	6.67	0.00	26.67	6.67	0.00	13.33	0.00	6.67	Negative
196	0.00	0.00	20.00	13.33	0.00	6.67	0.00	6.67	6.67	20.00	0.00	0.00	0.00	0.00	6.67	6.67	0.00	6.67	0.00	6.67	Negative
197	0.00	0.00	0.00	0.00	6.67	0.00	0.00	6.67	0.00	20.00	6.67	6.67	0.00	6.67	0.00	20.00	6.67	13.33	0.00	6.67	Negative
198	0.00	0.00	0.00	20.00	6.67	20.00	0.00	0.00	0.00	6.67	6.67	0.00	6.67	6.67	6.67	0.00	0.00	6.67	0.00	13.33	Negative
199	6.67	0.00	6.67	0.00	6.67	0.00	0.00	13.33	6.67	26.67	0.00	6.67	0.00	0.00	0.00	13.33	6.67	0.00	0.00	6.67	Negative

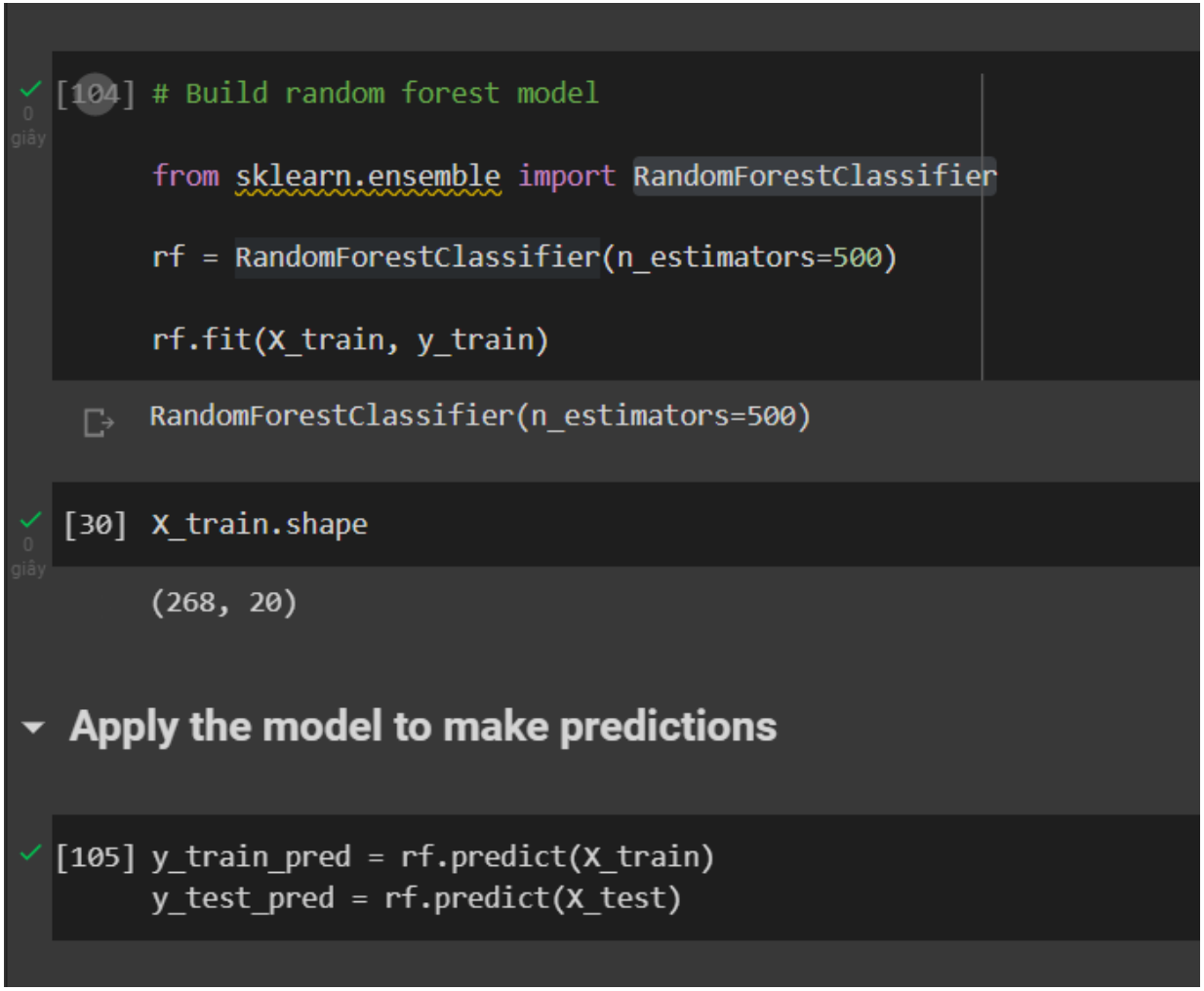
468 rows × 21 columns

```
# Encoding the Y class label
y = y.map({"Antiangio": 1, "Negative": 0})

y
```

```
0      1
1      1
2      1
3      1
4      1
      ..
195    0
196    0
197    0
198    0
199    0
Name: Class, Length: 468, dtype: int64
```

Thực hiện hệ thống

- Huấn luyện mô hình và kiểm tra kết quả dự đoán trên tập test :



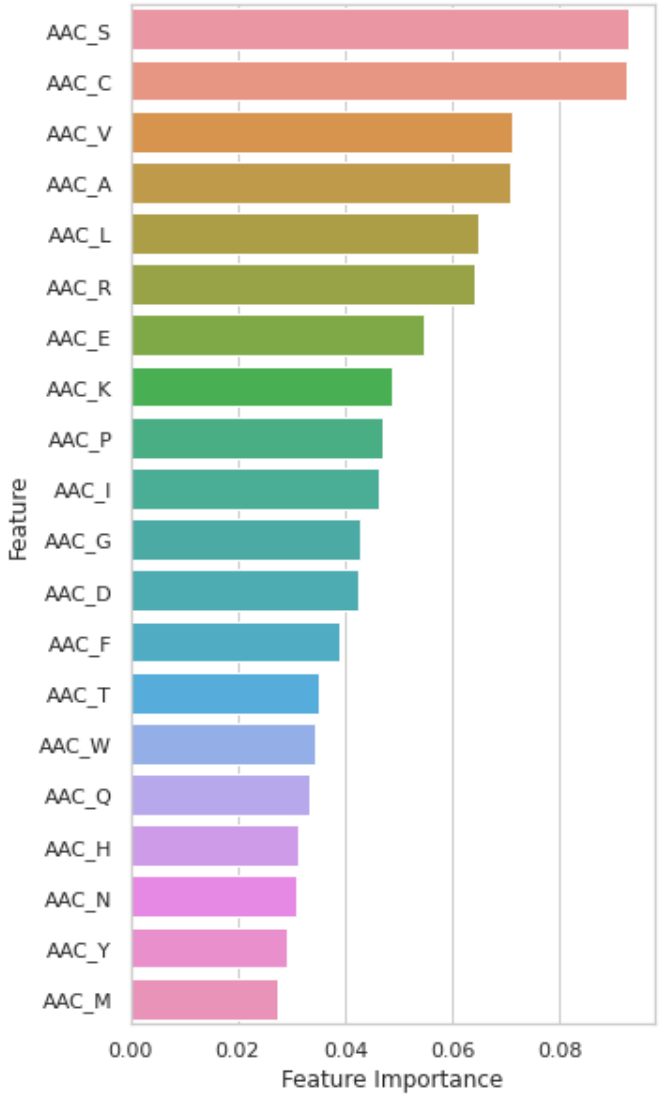
Các kết quả :

- Kết quả đạt được:
- Độ chính xác khá cao : 86%
- Chỉ số Sn, Sp cao và xấp xỉ nhau : 0.85,0.86
- Cho thấy tỉ lệ dự đoán đúng peptides chống tạo mạch và dự đoán đúng peptides không chống tạo mạch khá cao

	method	AC	Sn	Sp	MCC
RandomForestClassifier		0.860	0.854369	0.865979	0.720060

Các kết quả :

- Những đặc trưng AAC quan trọng trong việc dự đoán peptide chống tạo mạch



HUST

hust.edu.vn fb.com/dhbkhn

5.Các thử nghiệm khác :

Các thử nghiệm

- ✓ Thử nghiệm xây dựng mô hình với hơn 30 thuật toán học máy, đưa ra mô hình kết quả tốt nhất trên tập train

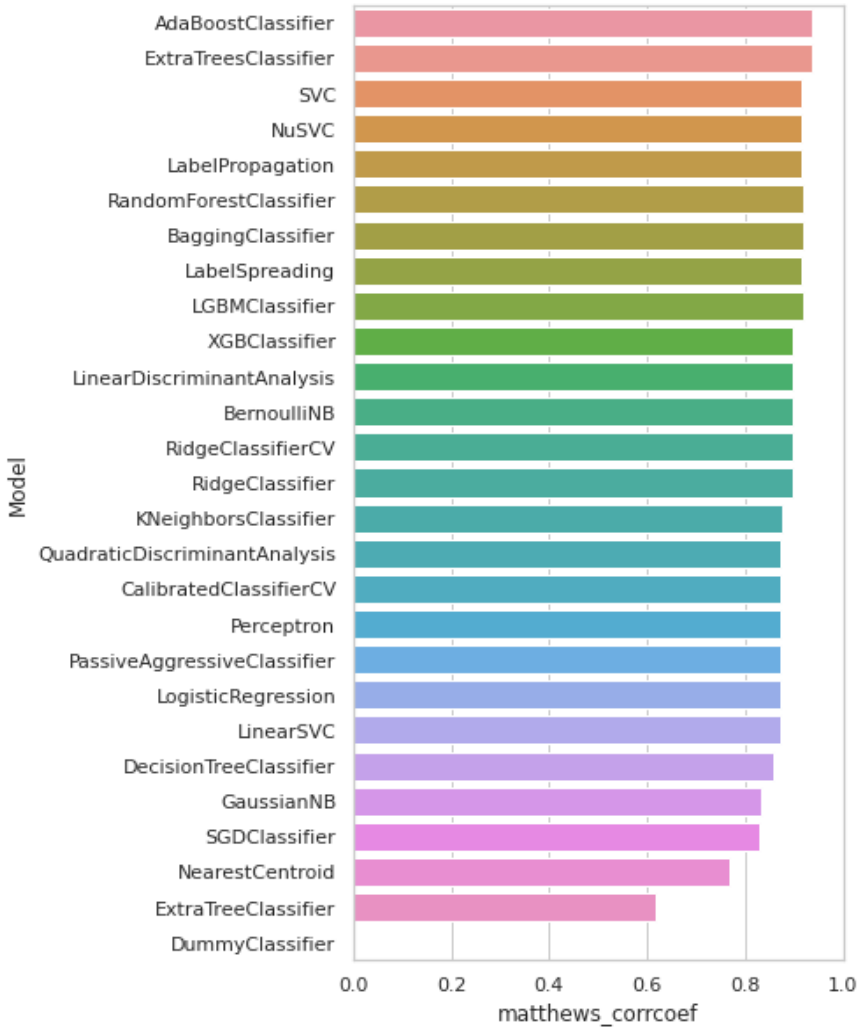
```
[ ]  feature = pd.read_csv('/content/featureall.csv')
```

```
[23]  # Import libraries
      import lazypredict
      from lazypredict.Supervised import LazyClassifier
      from sklearn.model_selection import train_test_split
      from sklearn.metrics import matthews_corrcoef

      # Load dataset
      X = feature.drop('Class', axis=1)
      y = feature['Class'].copy()

      y = y.map({"Antiangio": 1, "Negative": 0})
      #ytrain = ytrain.map({"Antiangio": 1, "Negative": 0})
      #ytest = ytest.map({"Antiangio": 1, "Negative": 0})


      # Data split
      X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state =42, stratify=y)

      # Defines and builds the lazyclassifier
      clf = LazyClassifier(verbose=0,ignore_warnings=True, custom_metric=matthews_corrcoef)
      #models_train,predictions_train = clf.fit(X_train, X_train, y_train, y_train)
      models_test,predictions_test = clf.fit(X_train, X_test, y_train, y_test)

      100%|██████████| 29/29 [00:02<00:00, 14.37it/s]
```

Các kết quả :

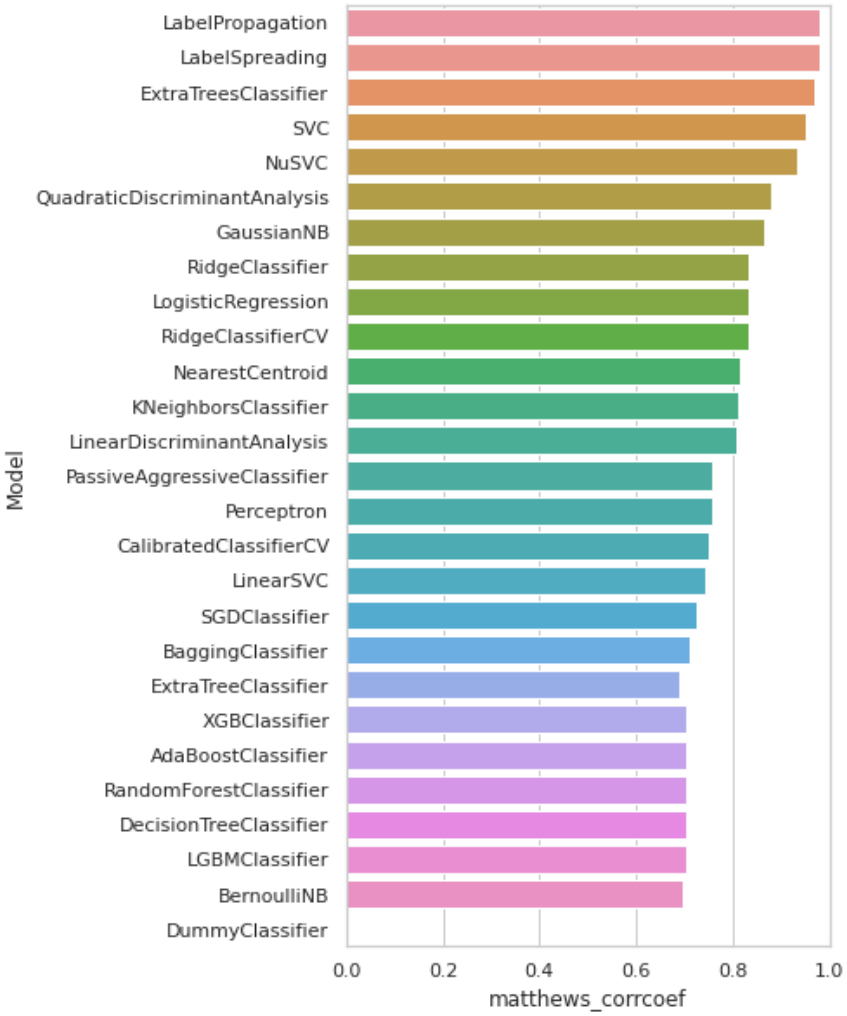
29 mô hình với tập train và tập test được lấy ra từ tập dữ liệu kết hợp của benchmark và NT15 với tỷ lệ 8-2

ta thấy 2 giải thuật như phân lớp adaboost, extratrees có độ chính xác cao nhất.



Các kết quả :

Tuy nhiên khi sử dụng tập benchmark làm tập train và NT15 làm tập test thì giải thuật phân lớp adaboost cũng như random forest giảm độ chính xác đáng kể, còn phân lớp extratrees vẫn khá cao.



HUST

hust.edu.vn fb.com/dhbkhn

- Do đó em thử nghiệm huấn luyện thêm 2 mô hình sử dụng 2 giải thuật phân lớp trên và kiểm tra kết quả dự đoán trên tập test. Tương tự cách huấn luyện mô hình random forest.
- với tập train và tập test được lấy ra bằng 2 cách

Các kết quả :

- So sánh các độ đo của 3 mô hình với tập train là tập benmark còn tập test là NT15

	method	AC	Sn	Sp	MCC
0	RandomForestClassifier	0.860	0.854369	0.865979	0.720060
1	ExtraTreesClassifier	0.925	0.898148	0.956522	0.851973
2	AdaBoostClassifier	0.775	0.780000	0.770000	0.550028

- So sánh các độ đo của 3 mô hình với tập train và tập test được lấy ra từ tập dữ liệu kết hợp của benchmark và NT15 với tỷ lệ 8-2

	method	AC	Sn	Sp	MCC
0	RandomForestClassifier	0.797872	0.804348	0.791667	0.595880
1	ExtraTreesClassifier	0.872340	0.843137	0.906977	0.747392
2	AdaBoostClassifier	0.691489	0.687500	0.695652	0.383065

Các kết quả :

- Kết quả
  Xây dựng 3 mô hình và kiểm tra kết quả dự đoán trên tập test khi sử dụng tập train và tập test được lấy ra từ tập dữ liệu kết hợp của benchmark và NT15 với tỷ lệ 8-2 (lấy ngẫu nhiên) với :

- Đặc trưng DPC

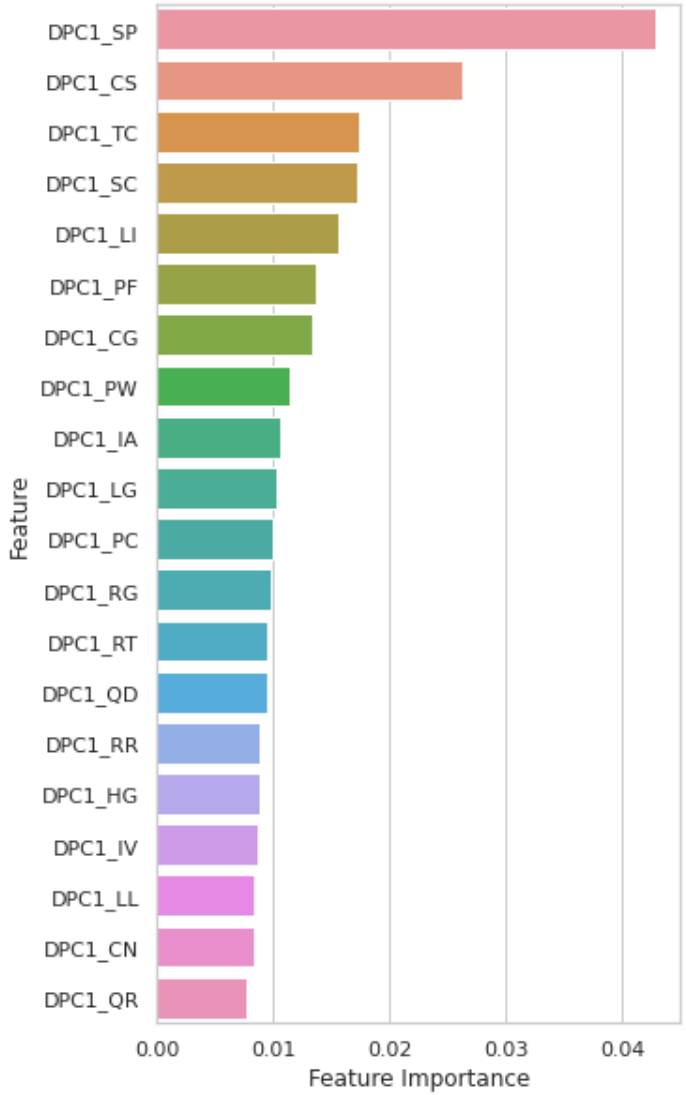
	method	AC	Sn	Sp	MCC
0	RandomForestClassifier	0.755319	0.772727	0.740000	0.511682
1	ExtraTreesClassifier	0.840426	0.847826	0.833333	0.681005
2	AdaBoostClassifier	0.734043	0.761905	0.711538	0.470757

- Đặc trưng AAC

	method	AC	Sn	Sp	MCC
0	RandomForestClassifier	0.797872	0.804348	0.791667	0.595880
1	ExtraTreesClassifier	0.872340	0.843137	0.906977	0.747392
2	AdaBoostClassifier	0.691489	0.687500	0.695652	0.383065

Các kết quả :

- Những đặc trưng DPC quan trọng trong việc dự đoán peptide chống tạo mạch



Thảo luận

Nhóm xin tự đánh giá những thứ mà chúng em tự thu nhận và đóng góp được cho bản thân sau đề tài:

- ✓Ứng dụng được kiến thức môn Tin sinh học vào một đề tài thực tiễn.
- ✓Cải thiện kiến thức, kỹ năng về mảng học máy và trí tuệ nhân tạo.
- ✓Rèn luyện kỹ năng làm việc nhóm.
- ✓Rèn luyện kỹ năng nghiên cứu và viết một báo cáo theo hướng nghiên cứu khoa học.

Tổng kết và phương hướng phát triển

- ✓ Nhóm đã hoàn thành mục tiêu đề ra: xây dựng một mô hình học máy dự đoán peptide chống tạo mạch từ đó hiểu hơn về môn Tin sinh học và những ứng dụng thực tiễn của nó trong cuộc sống
- ✓Về phương hướng phát triển trong tương lai: tìm cách cải thiện mô hình bằng cách sử dụng tập dữ liệu lớn hơn, thử xây dựng mô hình bằng phương pháp Deep learning.

HUST

THANK YOU !