RateMyAntibody

By:-
Samad Shahid, Shashwat Goyal, Srijan Garg,
Vaibhav Soni, Abhimanyu Lakra, Abhishek Soni

Group No: 5

Problem Statement

Prediction of antibody sequence effectiveness and its neutralising abilities against covid-19 and how it can further be used for effective diagnosis of patients and improved drug manufacturing.

Background Rationale

As we know Antibodies are the main line of defense against pathogens in human body. According to ([1]) The exposure to a pathogen produces B cells than can recognise and neutralize the pathogen. However not all antibodies are able to effectively bind/neutralise the antigen and hence they can’t prevent the disease from spreading. This goes to show the importance for the study of antibodies in order to correctly recognize and experiment upon a disease.

Source ([2]) performs comprehensive analysis of antibody response in 229 samples collected frequently from hospitalized COVID-19 patients. The result of this experiment shows that The majority of the hospitalized patients developed hACE2-blocking antibodies as well as neutralizing antibodies.

The findings from these sources motivated our project in which we will analyze data of antibodies received from recovered/expired Covid-19 patients and create a model which will predict antibody sequence and its binding and neutralizing effectiveness.

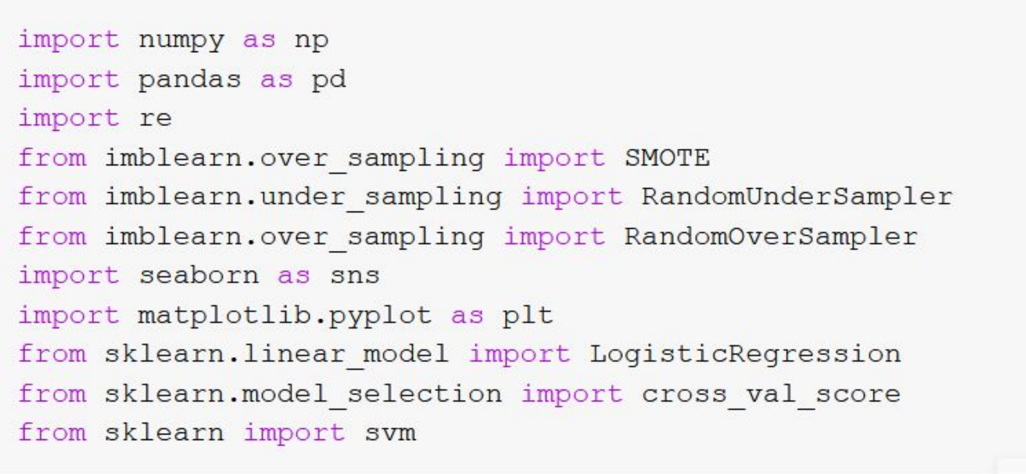
Methodology

1.) Data importing
2.) Data Preprocessing
3.) P-Feature generation & Amino acid Composition Feature
4.) 1st order dipeptide feature generation
5.) Machine Learning Model fitting

1) Data import

Importing Libraries

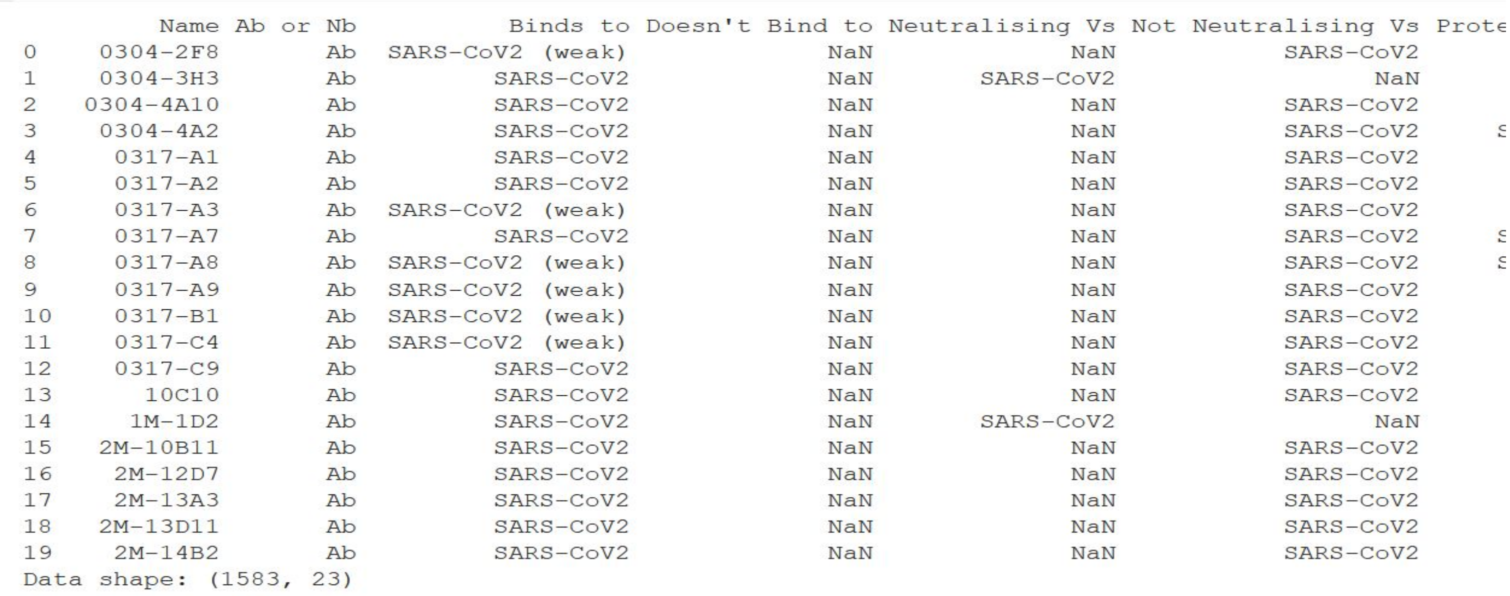
We have imported all the necessary libraries like numpy, pandas, sklearn, etc.



```
import numpy as np
import pandas as pd
import re
from imblearn.over_sampling import SMOTE
from imblearn.under_sampling import RandomUnderSampler
from imblearn.over_sampling import RandomOverSampler
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import cross_val_score
from sklearn import svm
```

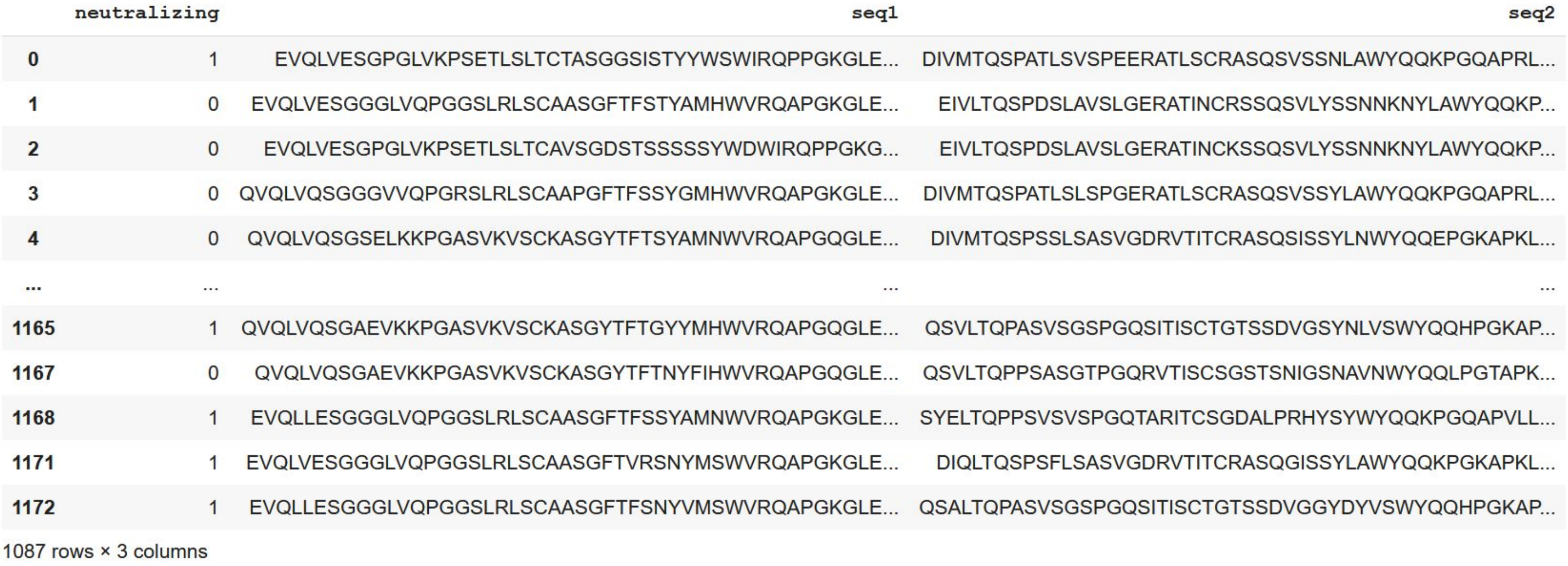
2) Dataset Preprocessing

Data was collected of various antibodies from http://opig.stats.ox.ac.uk/webapps/covabdab/ produced in response to COVID-19 from multiple patients.To ensure that these antibodies are antibodies associated with COVID-19, we have picked the antibodies binding to Sars-Cov-2. Finally we obtained 1583 rows with 23 columns.



```
         Name Ab or Nb          Binds to Doesn't Bind to Neutralising Vs Not Neutralising Vs Prote
0    0304-2F8       Ab  SARS-CoV2 (weak)             NaN             NaN           SARS-CoV2
1    0304-3H3       Ab         SARS-CoV2             NaN       SARS-CoV2                 NaN
2   0304-4A10       Ab         SARS-CoV2             NaN             NaN           SARS-CoV2
3    0304-4A2       Ab         SARS-CoV2             NaN             NaN           SARS-CoV2     S
4     0317-A1       Ab         SARS-CoV2             NaN             NaN           SARS-CoV2
5     0317-A2       Ab         SARS-CoV2             NaN             NaN           SARS-CoV2
6     0317-A3       Ab  SARS-CoV2 (weak)             NaN             NaN           SARS-CoV2
7     0317-A7       Ab         SARS-CoV2             NaN             NaN           SARS-CoV2     S
8     0317-A8       Ab  SARS-CoV2 (weak)             NaN             NaN           SARS-CoV2     S
9     0317-A9       Ab  SARS-CoV2 (weak)             NaN             NaN           SARS-CoV2
10    0317-B1       Ab  SARS-CoV2 (weak)             NaN             NaN           SARS-CoV2
11    0317-C4       Ab  SARS-CoV2 (weak)             NaN             NaN           SARS-CoV2
12    0317-C9       Ab         SARS-CoV2             NaN             NaN           SARS-CoV2
13      10C10       Ab         SARS-CoV2             NaN             NaN           SARS-CoV2
14     1M-1D2       Ab         SARS-CoV2             NaN       SARS-CoV2                 NaN
15   2M-10B11       Ab         SARS-CoV2             NaN             NaN           SARS-CoV2
16    2M-12D7       Ab         SARS-CoV2             NaN             NaN           SARS-CoV2
17    2M-13A3       Ab         SARS-CoV2             NaN             NaN           SARS-CoV2
18   2M-13D11       Ab         SARS-CoV2             NaN             NaN           SARS-CoV2
19    2M-14B2       Ab         SARS-CoV2             NaN             NaN           SARS-CoV2
Data shape: (1583, 23)
```

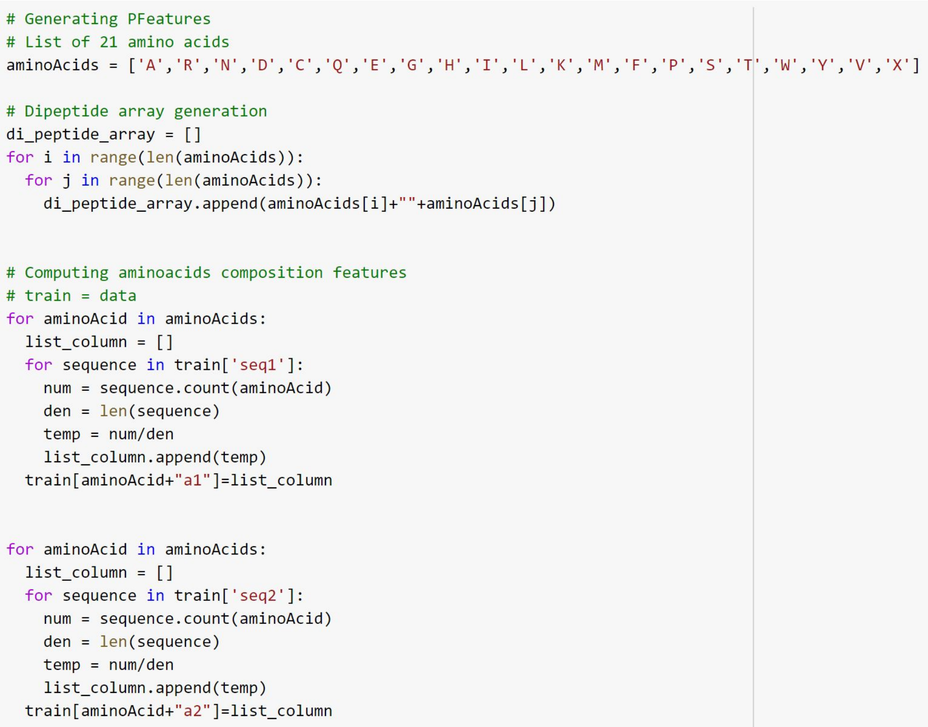
After removing the unnecessary columns, Neutralizing column only SARS-COV2 as 1 all else 0, removed the weak antibody rows, Removing rows containing ND values, and considering only VH or VHH and VL sequences, we get the following.



	neutralizing	seq1	seq2
0	1	EVQLVESGPGLVKPSETLSLTCTASGGSISTYYWSWIRQPPGKGLE...	DIVMTQSPATLSVSPEERATLSCRASQSVSSNLAWYQQKPGQAPRL...
1	0	EVQLVESGGGLVQPGGSLRLSCAASGFTFSTYAMHWVRQAPGKGLE...	EIVLTQSPDSLAVSLGERATINCRSSQSVLYSSNNKNYLAWYQQKP...
2	0	EVQLVESGPGLVKPSETLSLTCAVSGDSTSSSSSYWDWIRQPPGKG...	EIVLTQSPDSLAVSLGERATINCKSSQSVLYSSNNKNYLAWYQQKP...
3	0	QVQLVQSGGGVVQPGRSLRLSCAAPGFTFSSYGMHWVRQAPGKGLE...	DIVMTQSPATLSLSPGERATLSCRASQSVSSYLAWYQQKPGQAPRL...
4	0	QVQLVQSGSELKKPGASVKVSCKASGYTFTSYAMNWVRQAPGQGLE...	DIVMTQSPSSLSASVGDRVTITCRASQSISSYLNWYQQEPGKAPKL...
...	...	...	...
1165	1	QVQLVQSGAEVKKPGASVKVSCKASGYTFTGYYMHWVRQAPGQGLE...	QSVLTQPASVSGSPGQSITISCTGTSSDVGSYNLVSWYQQHPGKAP...
1167	0	QVQLVQSGAEVKKPGASVKVSCKASGYTFTNYFIHWVRQAPGQGLE...	QSVLTQPPSASGTPGQRVTISCSGSTSNIGSNAVNWYQQLPGTAPK...
1168	1	EVQLLESGGGLVQPGGSLRLSCAASGFTFSSYAMNWVRQAPGKGLE...	SYELTQPPSVSVSPGQTARITCSGDALPRHYSYWYQQKPGQAPVLL...
1171	1	EVQLVESGGGLVQPGGSLRLSCAASGFTVRSNYMSWVRQAPGKGLE...	DIQLTQSPSFLSASVGDRVTITCRASQGISSYLAWYQQKPGKAPKL...
1172	1	EVQLLESGGGLVQPGGSLRLSCAASGFTFSNYVMSWVRQAPGKGLE...	QSALTQPASVSGSPGQSITISCTGTSSDVGGYDYVSWYQQHPGKAP...

1087 rows × 3 columns

3) PFeature Generation & Amino acid Composition Feature

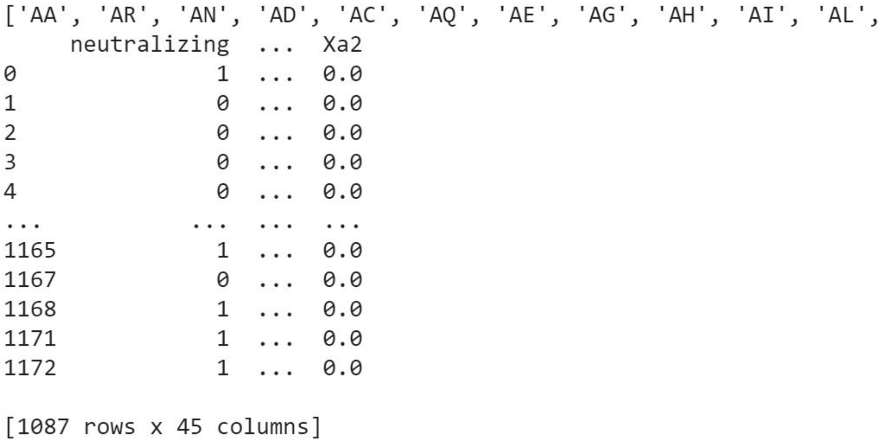


```python
# Generating PFeatures
# List of 21 amino acids
aminoAcids = ['A','R','N','D','C','Q','E','G','H','I','L','K','M','F','P','S','T','W','Y','V','X']

# Dipeptide array generation
di_peptide_array = []
for i in range(len(aminoAcids)):
  for j in range(len(aminoAcids)):
    di_peptide_array.append(aminoAcids[i]+""+aminoAcids[j])


# Computing aminoacids composition features
# train = data
for aminoAcid in aminoAcids:
  list_column = []
  for sequence in train['seq1']:
    num = sequence.count(aminoAcid)
    den = len(sequence)
    temp = num/den
    list_column.append(temp)
  train[aminoAcid+"a1"]=list_column


for aminoAcid in aminoAcids:
  list_column = []
  for sequence in train['seq2']:
    num = sequence.count(aminoAcid)
    den = len(sequence)
    temp = num/den
    list_column.append(temp)
  train[aminoAcid+"a2"]=list_column
```



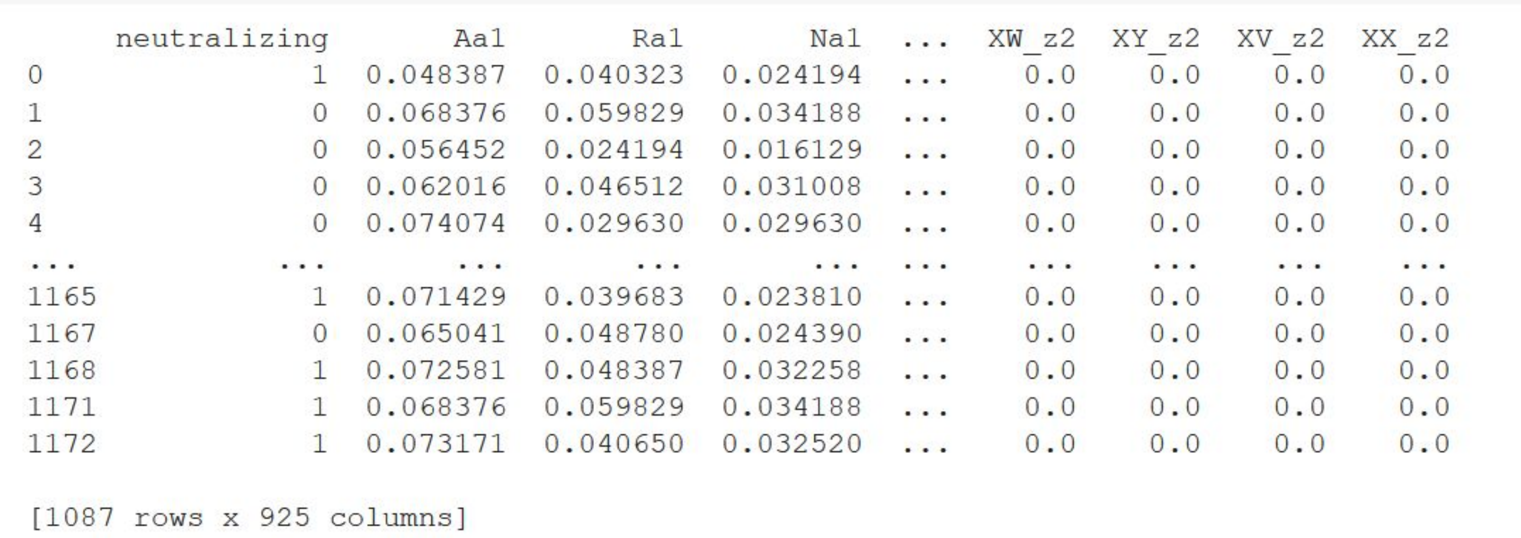
```
['AA', 'AR', 'AN', 'AD', 'AC', 'AQ', 'AE', 'AG', 'AH', 'AI', 'AL',
      neutralizing  ...  Xa2
0                1  ...  0.0
1                0  ...  0.0
2                0  ...  0.0
3                0  ...  0.0
4                0  ...  0.0
...            ...  ...  ...
1165             1  ...  0.0
1167             0  ...  0.0
1168             1  ...  0.0
1171             1  ...  0.0
1172             1  ...  0.0

[1087 rows x 45 columns]
```

We have considered 21 amino acids, generated feature columns using amino acids and dipeptides and done oversampling of the data to increase samples.

4)1st order dipeptide Feature Generation

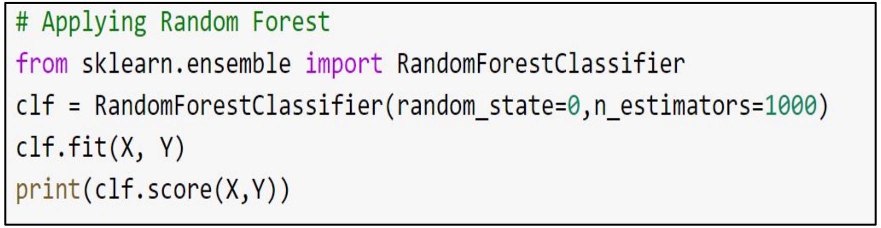
After applying the 1st order feature generation using dipeptide array, we get the the following matrix with multiple new features generated.



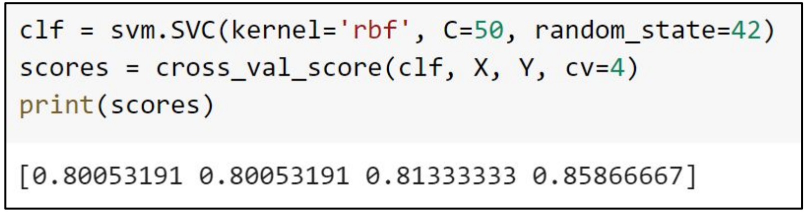
```
      neutralizing       Aa1       Ra1       Na1  ...  XW_z2  XY_z2  XV_z2  XX_z2
0                1  0.048387  0.040323  0.024194  ...    0.0    0.0    0.0    0.0
1                0  0.068376  0.059829  0.034188  ...    0.0    0.0    0.0    0.0
2                0  0.056452  0.024194  0.016129  ...    0.0    0.0    0.0    0.0
3                0  0.062016  0.046512  0.031008  ...    0.0    0.0    0.0    0.0
4                0  0.074074  0.029630  0.029630  ...    0.0    0.0    0.0    0.0
...            ...       ...       ...       ...  ...    ...    ...    ...    ...
1165             1  0.071429  0.039683  0.023810  ...    0.0    0.0    0.0    0.0
1167             0  0.065041  0.048780  0.024390  ...    0.0    0.0    0.0    0.0
1168             1  0.072581  0.048387  0.032258  ...    0.0    0.0    0.0    0.0
1171             1  0.068376  0.059829  0.034188  ...    0.0    0.0    0.0    0.0
1172             1  0.073171  0.040650  0.032520  ...    0.0    0.0    0.0    0.0

[1087 rows x 925 columns]
```

5)ML models applied

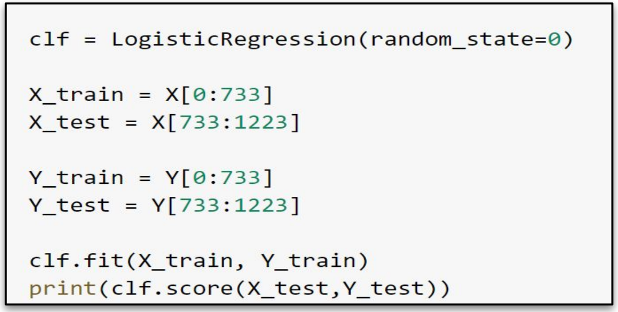


```
# Applying Random Forest
from sklearn.ensemble import RandomForestClassifier
clf = RandomForestClassifier(random_state=0,n_estimators=1000)
clf.fit(X, Y)
print(clf.score(X,Y))
```



```
clf = svm.SVC(kernel='rbf', C=50, random_state=42)
scores = cross_val_score(clf, X, Y, cv=4)
print(scores)
```

```
[0.80053191 0.80053191 0.81333333 0.85866667]
```



```
clf = LogisticRegression(random_state=0)

X_train = X[0:733]
X_test = X[733:1223]

Y_train = Y[0:733]
Y_test = Y[733:1223]

clf.fit(X_train, Y_train)
print(clf.score(X_test,Y_test))
```

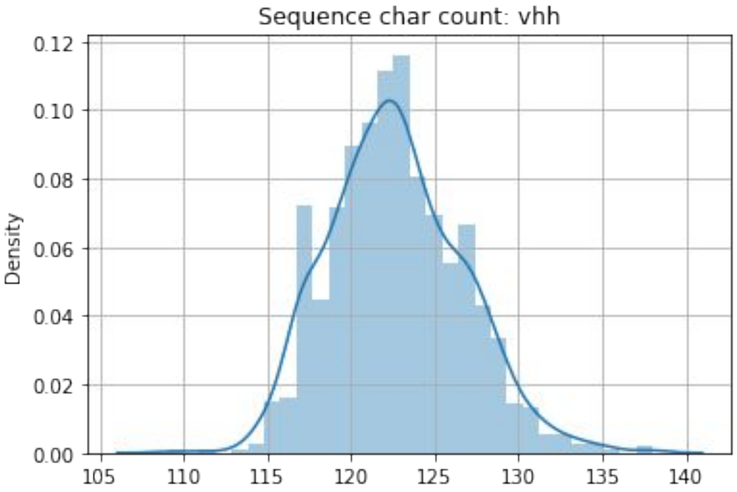
We used and evaluated our data using multiple machine learning models like Random Forest, Logistic regression, support vector machine and many more from the sklearn library. We used cross validation score to evaluate our data taking CV=4.

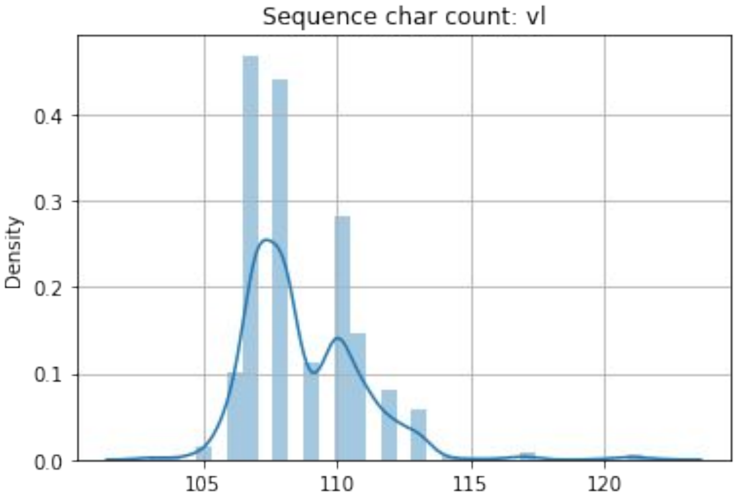
Results

- Our project outputs the prediction percentage of an antibody sequence ability to neutralize Sars-Covid-2.
- Prediction of whether antibodies for Covid-19 are able to neutralize the pathogen effectively or not and associate it with the clinical outcome of the patients. Difference between the effective and non-effective antibodies can help in better understanding of disease and development of efficacious vaccines. This can further be used for effective drug manufacturing.

Graphs

Sequence char count for the respective VHH and VL sequence.





Conclusion

From the sequence of the antibody of a person, we can determine whether their antibodies will effectively neutralize Covid-19 or not.

Our project could be applied on the ongoing trials for effectiveness of the monoclonal antibodies against the sars-cov2.

https://www.antibodysociety.org/covid-19-biologics-tracker/

https://chineseantibody.org/covid-19-track/

Contribution

Vaibhav Soni and **Samad Shahid** collected data and did exploratory data analysis

Abhishek Soni and **Abhimanyu Lakra** did feature generation and trained the Machine Learning Model

Srijan Garg and **Shashwat Goyal** did the preprocessing and generated gene features

Thank You!