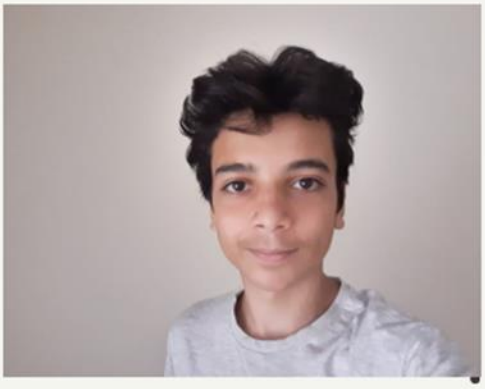
# 2022 年臺灣國際科學展覽會
# 優勝作品專輯

| | |
|---|---|
| 作品編號 | 080012 |
| 參展科別 | 生物化學 |
| 作品名稱 | BIOINFORMATIC PREDICTION OF CORONAVIRUS (SARS-COV-2) MUTATIONS THAT INCREASE CONTAGIOUSNESS |
| 得獎獎項 | |

| | |
|---|---|
| 國　　家 | Turkey |
| 就讀學校 | Buca Municipality Kızılçullu Science and Art Center-BUCA IMSEF |
| 指導教師 | Dr. Cansu ?lke Kuru |
| 作者姓名 | Murat Isik<br>Ege Caliskan |

關鍵詞　Coronavirus、COVID-19 pandemic、prediction algorithm

# 作者照片



Murat Işık

Ege Çalışkan

## Abstract

The COVID-19 pandemic, which emerged in 2019 and affected 223 countries, has caused nearly 260 million cases and 5.2 million deaths until today (November, 2021). The COVID-19 pandemic has reached this level of contagion as a result of the rapid spread of the SARS-CoV-2 virus and its advantageous variants. The aforementioned advantageous variants have occurred mainly through mutations seen on a single amino acid basis. These may cause changes in the structure of SARS-CoV-2, as well as impact the efficiency of interaction with the ACE2 protein, which the virus uses as the first step to enter the human host. Spike N501Y and E484K mutations that affect binding of spike with ACE2 have been widely observed in the UK, South Africa and Brazil since the beginning of 2020, and have caused concern all over the world. In the study, it was aimed to predict the SARS-CoV-2 mutations that could be as impactful as N501Y and E484K and could pose a danger due to their high contagiousness. To this end, experimental data on SARS-CoV-2 and ACE2 binding and stability were associated with different amino acid properties and integrated into a machine learning protocol, which showed that the N501M, Q414A, N354K, Q498H, N460K, N501W spike mutations are likely to have as dangerous effects as N501Y and E484K on the spread of the SARS-CoV-2 . At the end of the analyzes performed on the multi-interaction data we created with the ACE2 and RBD interaction data during the project development phase, the positions where dangerous variants can be seen were determined as G446, G447, Y505, T500, Q493, Y473 and G476. We suggest that particular attention should be paid to the mutations occurring on the 501st position, as we saw this position appearing repeatedly at the top of the "dangerous mutation" lists.

## Aim

Within the scope of our project, we aimed to predict which SARS-CoV-2 mutations will increase the infectiousness of the virus by using machine learning and computational biology techniques. For this purpose, we followed the steps below:

- ✓ Collecting information on mutations affecting the interaction between SARS-CoV-2 and its target protein ACE2 from the literature,
- ✓ Obtaining and processing amino acid change properties that will biochemically characterize the effect of SARS-CoV-2 mutations from the literature,
- ✓ To investigate which SARS-CoV-2 mutations may have the same effect as N501Y and E484K mutations, using k-mean and expectation maximization algorithms to reveal the pattern in the information listed above,
- ✓ Evaluation of mutations in the same class within the framework of the interaction between SARS-CoV-2 and ACE2.

## 1. Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is the virus responsible for the global pandemic that continues to infect millions of people worldwide as of February 2020 (Rodrigues et al., 2020). SARS-CoV-2 is the seventh coronavirus to be identified as a human pathogen. Among these coronaviruses, HKU1, NL63, OC43 and 229E types can cause mild symptoms as a result of human transmission, while SARS-CoV, MERS-CoV and SARS-CoV-2 viruses cause serious diseases in humans. Coronavirus 2019 (COVID-19) disease caused by SARS-CoV-2 is currently the most important public health problem in the world (Verma et al., 2020). COVID-19 infection commonly manifests as pneumonia, fever and dry cough. As a result of an excessive immune system response in some of the people with COVID-19 infection, serious damage to vital organs such as the lungs, heart, liver and kidneys occurs (Rajgor et al., 2020; Bhaskar et al., 2020; Valizadeh et al. 2020). SARS-CoV-2, whose contagiousness is very high compared to SARS-CoV and MERS-CoV, spread rapidly all over the world from Wuhan (China) since the end of 2019 and started to show its effect in our country as of Spring 2020 (Cucinotta and Vanelli , 2020; Jiang et al., 2019; Mahase et al., 2020). COVID-19 disease has gone down in history as the biggest epidemic of the last 100 years, causing 260 million cases and 5.2 million deaths to date (November 2021) (WHO, 2021).

### 1.1. How does SARS-CoV-2 cause infection?

SARS-CoV-2, belonging to the *Coronavirinae* subfamily in the *Coronaviridae* family of the order *Nidovirales*, has an enveloped and spherical structure, 150-160 nanometers in size (Figure 1-left top-). Positive single-stranded RNA carrying the genetic material of SARS-CoV-2 is preserved in a capsule consisting of nucleoprotein (Figure 1-left top- gray capsule). It is possible for SARS-CoV-2 to recognize its target in the human host with its S-protein (spike protein) (Figure 1 - left upper - red protrusions) (Kannan et al., 2020). SARS-CoV-2 initiates infection by binding to receptors on host cells via S-protein. The entry receptor for both SARS-CoV-2 and the original SARS-CoV-1 is the human cell surface protein Angiotensin Converting Enzyme 2 (ACE2) (Figure 1 - middle part -). ACE2 is a type I membrane (cell membrane) protein that can be found in most important organs such as lungs, arteries, heart and kidney. ACE2 is responsible for lowering blood pressure by forming angiotensin (1-7), a vasodilator. Strong interaction of the S protein of the SARS-CoV-2 virus and the catalytic region of the ACE2 enzyme is the first step in the formation of the disease (https://covid19.tubitak.gov.tr/bilimsel-arastirma-paylasim-platformu/laboratuvarda-tasarlanan-ace2-mutasyonlarinin-etkisini). Receptor Binding Domain (RBD) of the S-protein binds primarily to ACE2. Because of its role in viral entry, binding domain recognition module is an important determinant of interspecies transition and evolution. In addition, RBD is the target of the strongest anti-SARSCoV-2-neutralizing antibodies identified to date, and many promising vaccine candidates also use module as the sole antigen (Starr et al., 2020).

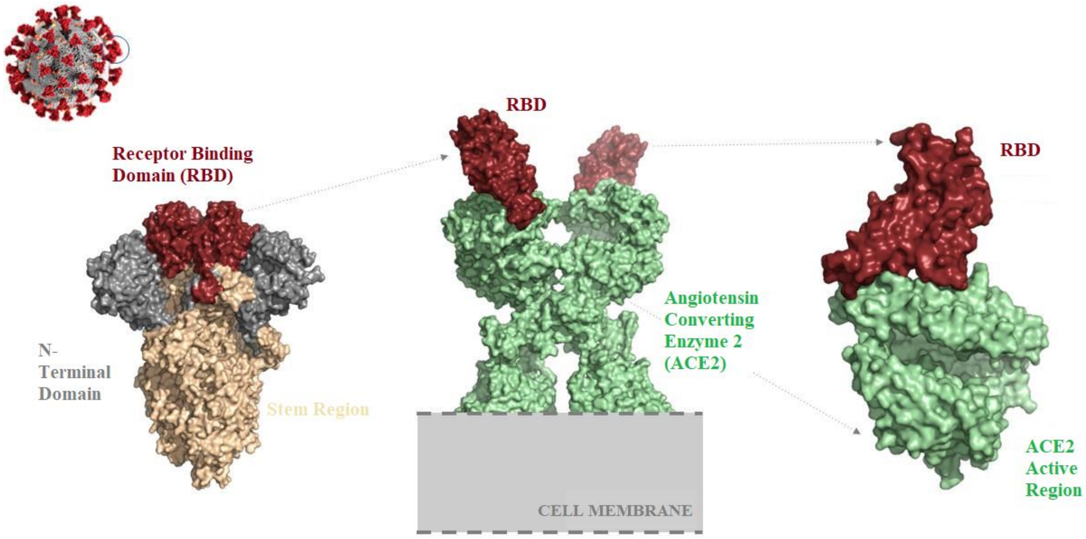

Receptor Binding Domain (RBD)
N-Terminal Domain
Stem Region
RBD
Angiotensin Converting Enzyme 2 (ACE2)
CELL MEMBRANE
RBD
ACE2 Active Region

**Figure 1. SARS-COV-2 and receptor angiotensin converting enzyme 2 (ACE2)** (https://covid19.tubitak.gov.tr/bilimsel-arastirma-paylasim-platformu/laboratuvarda-tasarlanan-ace2-mutasyonlarinin-etkisini). In this figure, the representative structure of SARS-CoV-2 is given at the top left. One of the S-proteins shown in red above this structure is shown in a circle. The S-protein consists of three regions: N-terminal Domain, Receptor Binding Domain (RBD) and Stem Region. RBD recognizes the ACE2 protein embedded in the cell membrane as shown in the middle and left of the figure. This stage is the first essential step for infection.

### 1.2. What is the relationship between SARS-CoV-2 and ACE2 interaction with virus mutation?

By binding ACE2 with SARS-CoV-2, the virus will have found the stable environment it needs to enter the cell. The virus then transmits the genetic RNA material into the cell, making the host cell produce a series of copies of itself. The newly formed viruses break down the cell, enter the blood and start the infection by selecting new target cells. Meanwhile, the immune system prepares an attack against virus spread and the antibodies it produces recognize primarily the S-proteins of the virus. S-proteins that shut down in this way, ie neutralized, can no longer bind to the ACE2 receptor (Starr et al., 2020). In this case, the body successfully overcomes the infection. However, the virus produces errors, that is, mutations, every time the virus replicates itself in a cell. Most of the mutations that occur are actually harmful to the virus, causing the virus to lose its effectiveness. These SARS-CoV-2 variants cannot survive because these mutations are at a disadvantage for the virus (Nelson et al., 2021). However, some types of coincidental mutations give the virus an advantage, and so the virus continues its task of finding new cells to infect more efficiently. Some of these advantageous mutations seen cause the S-protein to better adapt to ACE2. In this case, the rate of transmission of the virus is also increased. Since the beginning of 2020, the N501Y and E484K mutations, which are common in England, South Africa and Brazil, have caused the S-protein of the virus to bind more tightly to the aforementioned ACE2. These mutations are also thought to cause some antibodies that previously recognized S-proteins to no longer work (Luan et al., 2021). Therefore, the existence and understanding of these mutations is the basis for drug and vaccine studies.

This process is adapted from https://www.npr.org/sections/goatsandsoda/2021/02/02/961668700/whats-going-on-with-all-these-coronavirus-variants-an-illustrated-guide It is visually given in Figure 2.

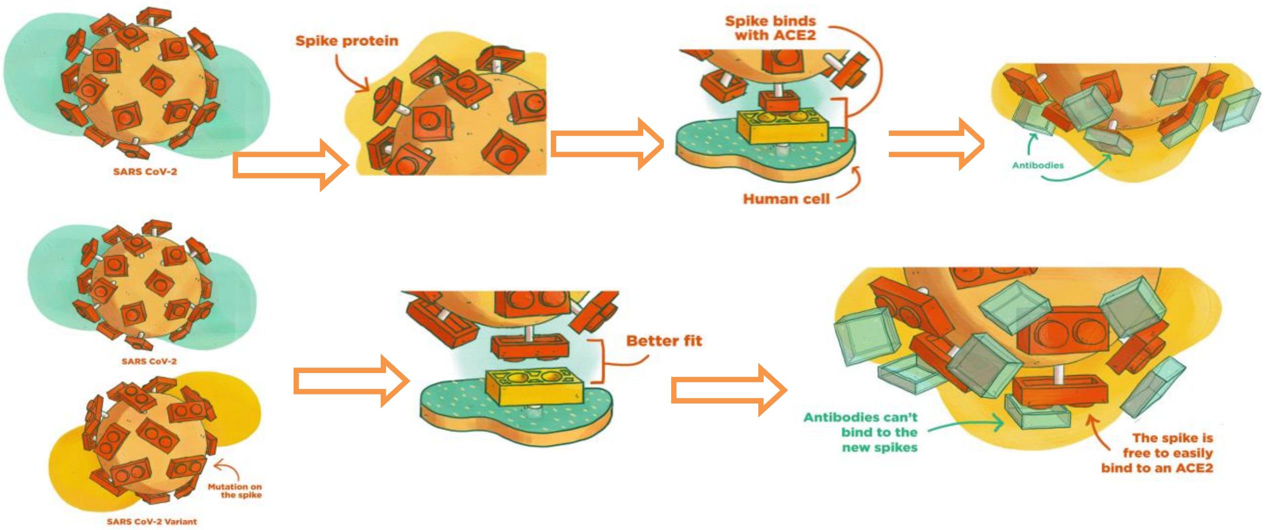


**Figure 2. The importance of binding of SARS-CoV-2 with ACE2 in the interpretation of infection and virus mutations** (https://www.npr.org/sections/goatsandsoda/2021/02/02/961668700/whats-going-on-with-all-these-coronavirus-variants-an-illustrated-guide In the first line, it was shown how antibodies prevent the binding of the spike (S) protein to ACE2 in the human cell, as it finds its matching puzzle partner. In the second line, it is depicted that the SARS-CoV-2 S-protein with the advantageous mutation binds better to the target and at the same time becomes unrecognizable by antibodies due to its altered structure.

### 1.3. Detected virus mutations that affect the interaction of SARS-CoV-2 with ACE2

The global economic recession caused by the COVID-19 pandemic is unusually severe, causing loss of livelihoods and income on a global scale. Under these conditions, a large number of scientists and researchers are making an unprecedented effort to find vaccines and therapeutics to halt the SARS-CoV-2 outbreak (Luan et al., 2021). As in other RNA viruses, the high mutation rate of the virus, which spreads rapidly, is one of the biggest obstacles to the sustainability of vaccines and drugs to be developed (van Dorp et al., 2020; Pachetti et al., 2020; Phan, 2020). In this context, it is of great importance to enlighten the structural and functional features of the genome of SARS-CoV-2 with all its aspects. The B.1.1.7 SARS-CoV-2 variant, which has emerged in the UK since the beginning of 2020, has made the virus 30% -50% more contagious and widespread. B.1.1.7 is a change in position 501 in the S-protein to which many neutralizing antibodies bind (Gu et al., 2020). Position 501, which was originally asparagine, was converted to the amino acid tyrosine as a result of this mutation (N501Y). This position is in RBD and improves the binding of the S-protein to ACE2. The molecular mechanism of this improved binding is still unclear, requiring evaluation of their effects on existing therapeutic antibodies (Luan et al., 2021). Again, at the beginning of 2020, although we do not yet have as much information about B.1.351, a second SARS-CoV-2 variant that emerged in South Africa and Brazil, this mutation is also known as the 484. is known to be in position. 484, originally glutamic acid, converted to lysine after mutation (E484K). Since its position is in RBD, it is thought that E484K can resist antibodies produced in people with COVID-19 (Nelson et al., 2021).

### 1.4. Is it possible to predict SARS-CoV-2 mutations that will increase contagiousness?

In September 2020, a seminal article on SARS-CoV-2 was published by Starr et al. (Starr et al., 2020). In this article, all possible mutations on the RBD region of the S-protein of SARS-CoV-2 were scanned and how these mutations affect the binding of SARS-CoV-2 to ACE2. According to this study, which screened 4221 S-protein RBD mutations, approximately 14% of these mutations improved the binding of the S-protein to ACE2. N501Y and E484K mutations, variants of England, Brazil and South Africa, are also included in this group of 14%. While there are 586 (14%) different mutations that can improve the binding of S-protein and ACE2, the fact that only two mutations are common yet indicates that not every mutation that may be advantageous for the virus has an equal chance. Within the scope of the project, it was aimed to find SARS-CoV-2 mutations that would have an advantage similar to the N501Y and E484K mutations, taking into account the biochemical changes caused by the mutation with the binding and S-protein stability data of Starr et al. In this context, it was aimed to introduce the experimental and literature information to different machine learning grouping algorithms, to find out which mutations would fall into the same class as the N501Y and E484K mutations, and thus to predict the contagiousness levels of future mutations.

## 2. Material-Method

### 2.1. Project Flowchart

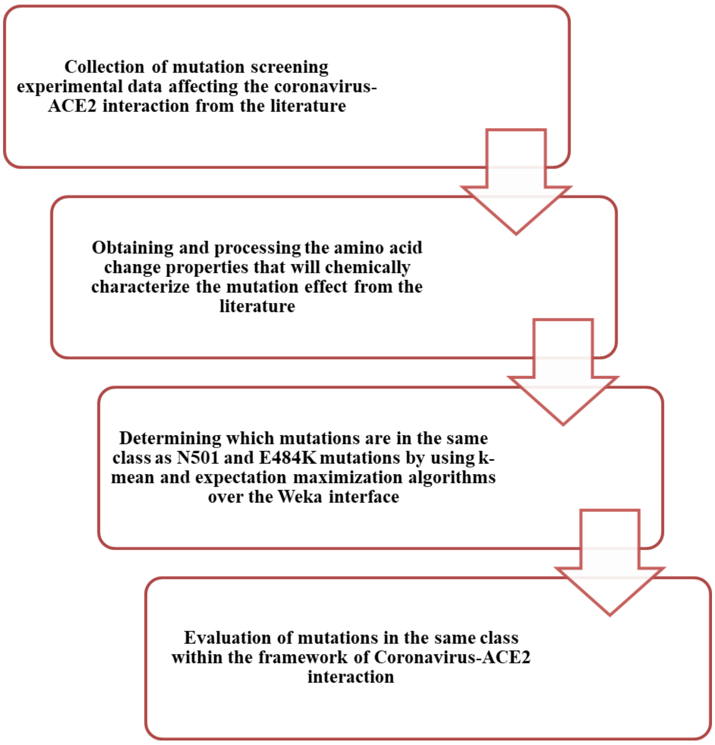

Collection of mutation screening experimental data affecting the coronavirus-ACE2 interaction from the literature
Obtaining and processing the amino acid change properties that will chemically characterize the mutation effect from the literature
Determining which mutations are in the same class as N501 and E484K mutations by using k-mean and expectation maximization algorithms over the Weka interface
Evaluation of mutations in the same class within the framework of Coronavirus-ACE2 interaction

**Figure 3.** Project Flowchart

### 2.2. Experimental data sets used in the project

Experimental data sets used within the scope of the project are given under the link as whole and sub-series. https://github.com/BiyoinformatikProje/Prediction-of-SARS-CoV-2-Mutations-that-Increase-Contagiousness

Experimental data sets were obtained from the article within the scope of the study performed by Starr et al. In 2020 (Starr et al., 2020 and https://jbloomlab.github.io/SARS-CoV-2-RBD_DMS ). The original of this set screened the effect of 4221 S-protein mutations on the stability of the S-protein and its binding to ACE2. Among these mutations, 586 either did not affect or improve the binding of the S-protein to ACE2. In this study, we focused on this subset, which corresponds to approximately 14% of the main set. For this subset, we got mutation information and expr_avg and bind_avg information from the data set linked above. While expr_avg provides information about the stability of the protein (the bigger the better), bind_avg measures how strongly the S-protein is bound to ACE2 (the bigger the better).

### 2.3. Amino acid properties used in the project

All proteins, whether from the oldest bacterial species or the most complex life forms, consist of 20 amino acid clusters that are covalently linked in characteristic linear sequences. Since each of these amino acids has a side chain with its own unique chemical properties, these 20 precursor molecule groups can be regarded as the alphabet in which the language of the protein structure is written. All 20 common amino acids are α-amino acids. They have a carboxyl group and an amino group attached to the same α-carbon atom. They differ from each other in their side chains or R groups, which vary in structure, size and electrical charge and affect the solubility of amino acids in water. Amino acids can be simplified by grouping them into five main classes based on the properties of the R groups, particularly their polarity or tendency to interact with water at biological pH (close to pH 7.0). The polarity of the R groups varies widely, from nonpolar and hydrophobic (insoluble in water) to highly polar and hydrophilic (water soluble). Their different solubilities are used to separate polypeptides. The water solubility of large polypeptides depends on the polarity of the R groups, especially the number of ionized groups: the more ionized groups there are, the more soluble the polypeptide (Lehninger et al., 2005). Within the scope of the project, we analyzed the changes in the amino acid mutations for the S-protein Receptor Binding Domain (RBD) in wild and mutant-type constructs on the basis of 8 different characteristics, based on the experimental data set we obtained from the article by Starr et al. These amino acid properties were determined as a result of the literature research as Hydropathy, Polarity, Volume, Molecular weight, ring number in amino acid structure, oxygen number, hydrogen number and double bond number within the scope of their effects on the average bonding and stability average.

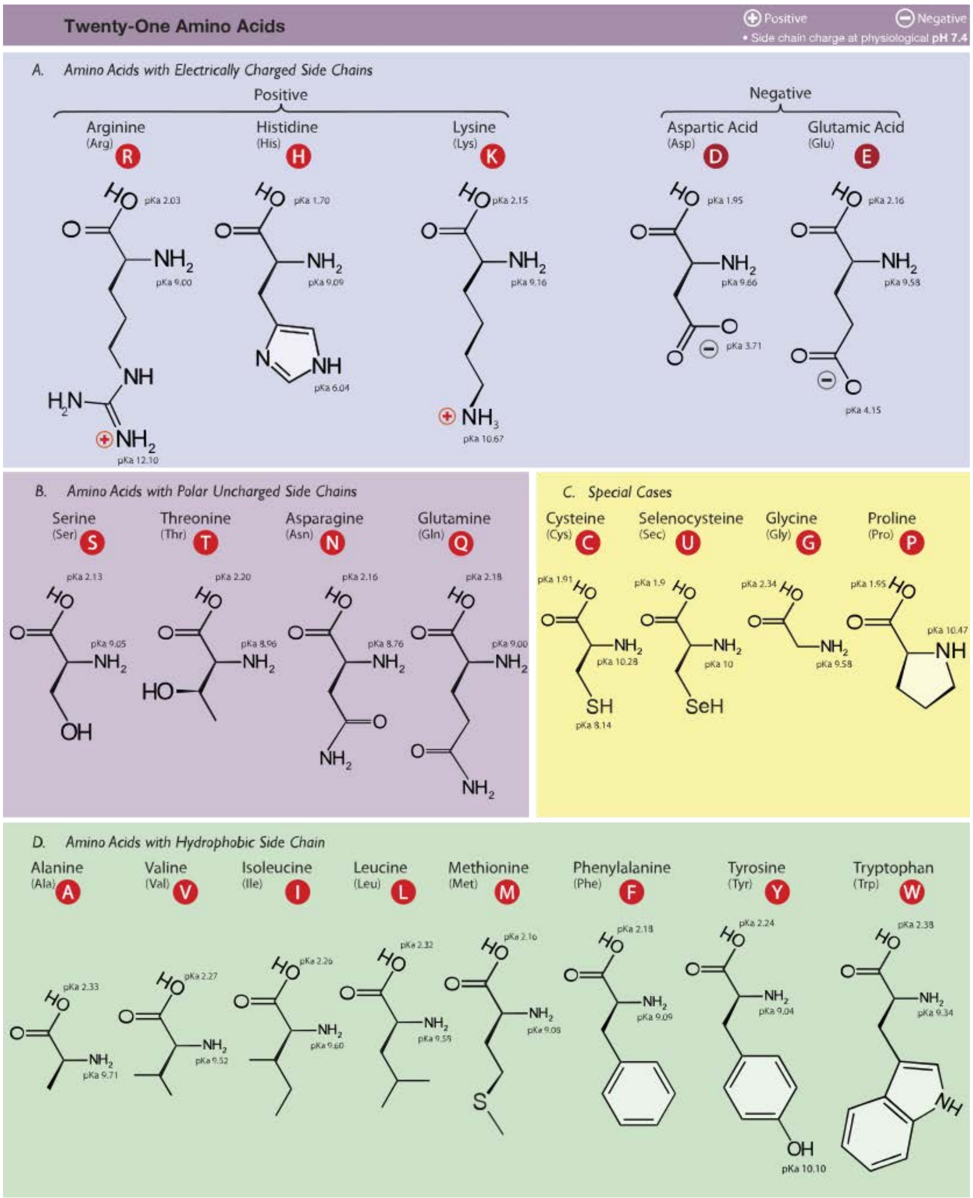


**Figure 4.** Amino acid features and classification

**Table 1.** Amino acid features used in the project

| Aminoacid features | Definition | Purpose of Use in the Project | References |
|---|---|---|---|
| **Volume** | It is the amount of space a substance or object occupies. | Classification based on determining the volume change, especially in side chains, due to mutation | (Amengual-Rigo et al., 2020). |
| **Hydropathy** | Interpreted as water repellency, the hydropathy index is a scale that reports the hydrophobic and hydrophilic tendency of a chemical group. Hydropathy index positive value shows high water repellency, ie hydrophobic feature. | Classification based on determining the change in the hydrophobicity of the side chains due to mutation. | (Amengual-Rigo et al., 2020). |
| **Moleculer Mass** | It is the value calculated by taking the sum of the atomic masses of the atoms in a molecule. | Classification based on determining the change in molecular mass of amino acids due to mutation | - |

| | | | |
|---|---|---|---|
| **Polarity** | It is the separation of electric charge that leads to a molecule with a negatively charged end and a positively charged end, or a chemical group with an electric dipole moment. | Classification based on determining the change in the charge state of amino acids due to mutation | (Sun et al., 2017) |
| **The number of cyclic structures in the amino acid structure** | It indicates the number of cyclic structures formed by bonding together the carbon atoms in the amino acid structure. | Classification based on determination of changes in polarity and hydropathy properties due to the aromatic ring structure and number in the R-group of the mutation-induced amino acid structure | - |
| **Oxygen number in amino acid structure** | It indicates the number of oxygen atoms in the amino acid structure. | Classification based on the difference in the number of oxygen atoms due to the change in the amino acid structure due to mutation | - |
| **Hydrogen number in amino acid structure** | It indicates the number of hydrogen atoms in the amino acid structure. | Classification based on the difference in the number of hydrogen atoms due to a change in the amino acid structure due to mutation<br>This property can also be associated with the number of cyclic structures. Since the carbon atoms at the ends are bonded to each other in ring structures, the number of hydrogen atoms the compound has naturally decreases. When the two ends of the compound are combined, one hydrogen atom will be removed from each end, so in cyclic structures with the same carbon number, two hydrogen atoms are missing compared to straight chains. | - |
| **The number of double bonds in the amino acid structure** | It indicates the number of double bonds in the amino acid structure. | Classification based on the difference in the number of double bonds due to change in amino acid structure due to mutation | - |

The changes of the properties given in Table 1 for the mutations in the selected data set were calculated with the python codes we wrote. Ultimately, the mutation and amino acid change information was put into a table using python, again as a single csv table. The grouping of the obtained table was provided with the Weka program.

## 2.4. Programs and algorithms

### 2.4.1. Machine Learning Data Analysis Program, WEKA

In the project, the machine learning program WEKA (Version 3.8.5.) Was used for grouping. WEKA, short for Waikato Environment for Knowledge Analysis, is a program that can be used for data analysis, analysis and predicting. In this project, the program was used to observe the distribution of the data and extract meaning and information from the data. K-means clustering and Expectation Maximization algorithms were used within the scope of this program. (https://www.tutorialspoint.com/weka/index.htm).

### 2.4.2. K-means Clustering

K-means clustering, which is an algorithm within the Weka data analysis program, is an unsupervised machine learning algorithm that works with the "k" variable determined by the user (Figures 4, 5). The variable "k" represents the number of sets needed before starting the algorithm. This clustering method divides a data set consisting of N data objects into k sets given as input parameters. The aim here is to ensure that the intra-cluster similarity of the clusters obtained at the end of the partitioning process is maximum and the similarity between clusters is minimum. K-average is one of the most commonly used and easy to implement clustering algorithms. This algorithm can cluster large-scale data quickly and effectively (Alpaydin, 2004).

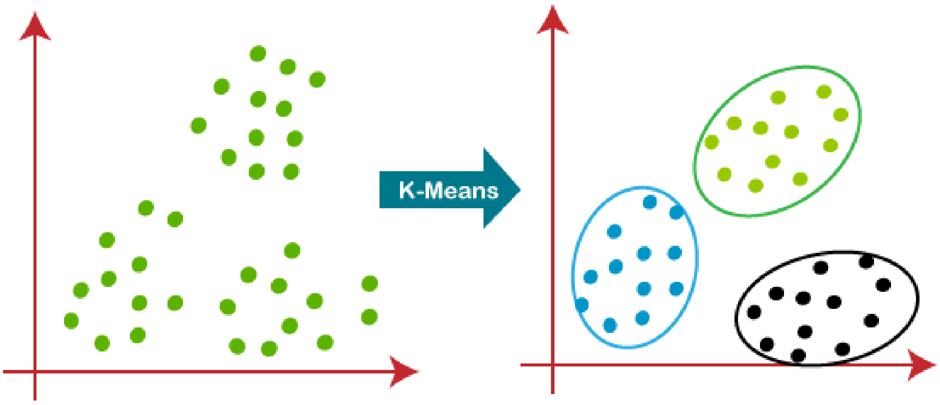
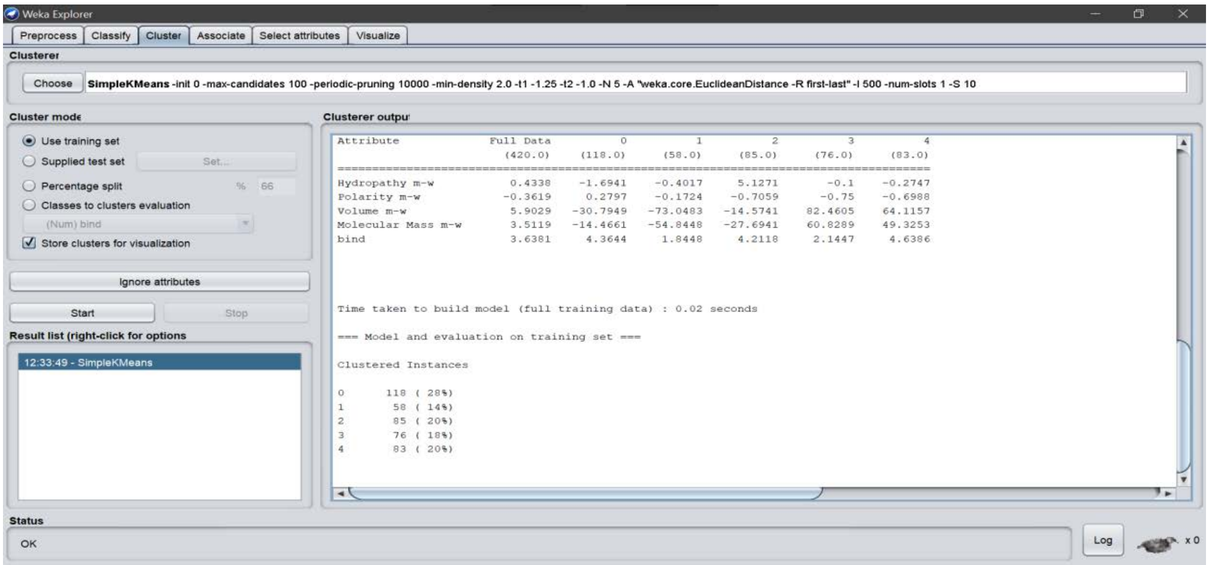


**Figure 5.** The K-mean clustering algorithm divides the data into k sets that are distinct from each other. (https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning)



SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 5 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10

- ● Use training set
- ○ Supplied test set
- ○ Percentage split % 66
- ○ Classes to clusters evaluation (Num) bind
- ☑ Store clusters for visualization

12:33:49 - SimpleKMeans

Clusterer outpu

```
Attribute                Full Data          0          1          2          3          4
                           (420.0)    (118.0)     (58.0)     (85.0)     (76.0)     (83.0)
=========================================================================================
Hydropathy m-w              0.4338    -1.6941    -0.4017     5.1271       -0.1    -0.2747
Polarity m-w               -0.3619     0.2797    -0.1724    -0.7059      -0.75    -0.6988
Volume m-w                  5.9029   -30.7949   -73.0483   -14.5741    82.4605    64.1157
Molecular Mass m-w          3.5119   -14.4661   -54.8448   -27.6941    60.8289    49.3253
bind                        3.6381     4.3644     1.8448     4.2118     2.1447     4.6386




Time taken to build model (full training data) : 0.02 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       118 ( 28%)
1        58 ( 14%)
2        85 ( 20%)
3        76 ( 18%)
4        83 ( 20%)
```

Status
OK

**Figure 6.** A visual that represents the result from Weka when using the K-mean algorithm.

### 2.4.3. Expectation-maximization (EM) Algorithm

This algorithm is an iterative search method used to find the greatest likelihood or the largest aftershock estimates of the parameters of statistical models based on unobservable hidden variables. It occurs by repeating two steps consecutively as the expectation (B) step and the maximization (M) step. The expectation step generates a log-likelihood expectation function using the current estimates of the parameters. The maximization step updates the parameter values to maximize the log-likelihood expectation. So each of these two steps feeds each other by calculating the other's input. Expectation maximization steps are repeated until the amount of error in the estimation falls below a certain rate (Figures 7, 8) (Alpaydin, 2020).

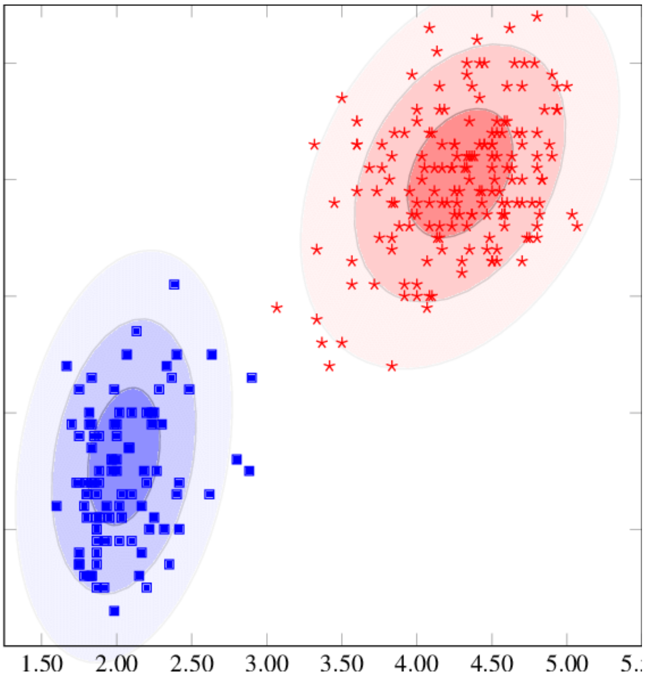
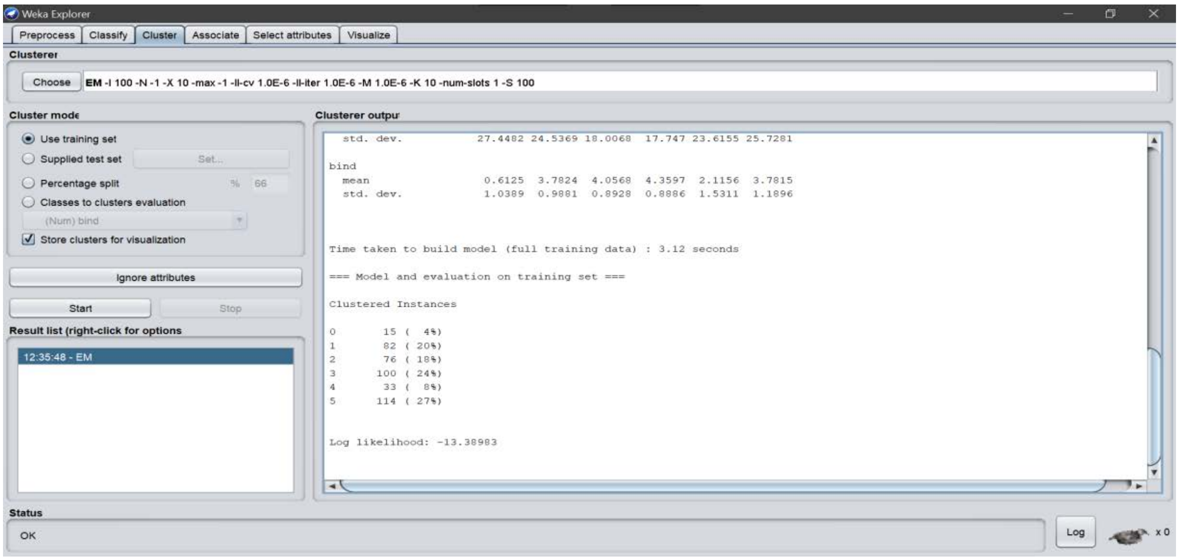


**Figure 7.** The Expectation Maximization algorithm separates the clusters at the maximum distance from each other (Stamp, adapted from 2018).



EM -I 100 -N -1 -X 10 -max -1 -ll-cv 1.0E-6 -ll-iter 1.0E-6 -M 1.0E-6 -K 10 -num-slots 1 -S 100

● Use training set
○ Supplied test set
○ Percentage split % 66
○ Classes to clusters evaluation
(Num) bind
☑ Store clusters for visualization

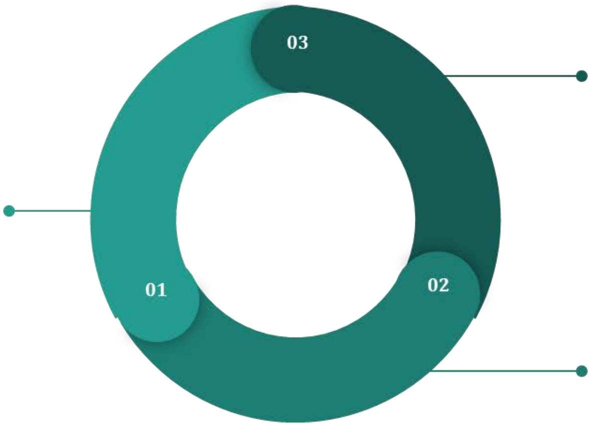
12:35:48 - EM

```
  std. dev.            27.4482 24.5369 18.0068  17.747 23.6155 25.7281

bind
  mean                  0.6125  3.7824  4.0568  4.3597  2.1156  3.7815
  std. dev.             1.0389  0.9881  0.8928  0.8886  1.5311  1.1896



Time taken to build model (full training data) : 3.12 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       15 (  4%)
1       82 ( 20%)
2       76 ( 18%)
3      100 ( 24%)
4       33 (  8%)
5      114 ( 27%)


Log likelihood: -13.38983
```

OK

**Figure 8.** A visual that represents the result from Weka when using the EM algorithm.

## 2.5. Project Development Process



**SARS_CoV_2 - We Generated ACE2 Multiple Interaction Data.**

In our Python code, we converted the ACE2 and RBD single interaction data, which we created by calculating our own atomistic structure, into multiple interaction data. You can find details about this in our github page in the "Data_Edit.ipynb" file.

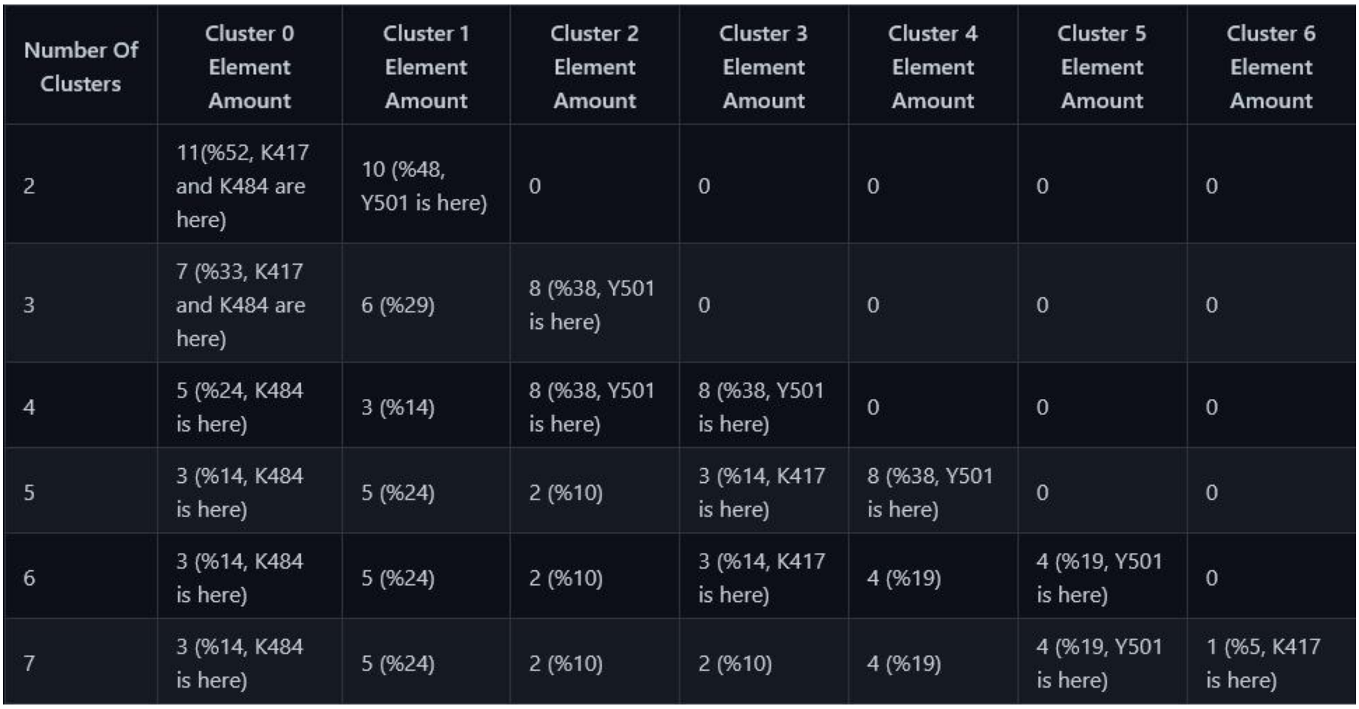
**We analyzed this data in WEKA Program and reached the conclusion.**

We tried various algorithms and cluster numbers on our new data in the Weka program. We determined the most optimal algorithm as the 6 Cluster K-Means algorithm. As a result of the analysis we made using this algorithm, we identified dangerous interactions.

**We've Updated the Description of Our Features to Comply with Our New Data.**

Our new data is based on interactions, not mutations, so it gives us a different perspective. Here, we have processed the properties as the sum of the properties of ACE2 amino acids with which RBD interacts. Details on this can be found on our github page in the "Data Editing.ipynb" file.

**Table 2.** K-Means Analysis Results in ACE2-RBD Dataset



| Number Of Clusters | Cluster 0 Element Amount | Cluster 1 Element Amount | Cluster 2 Element Amount | Cluster 3 Element Amount | Cluster 4 Element Amount | Cluster 5 Element Amount | Cluster 6 Element Amount |
|---|---|---|---|---|---|---|---|
| 2 | 11(%52, K417 and K484 are here) | 10 (%48, Y501 is here) | 0 | 0 | 0 | 0 | 0 |
| 3 | 7 (%33, K417 and K484 are here) | 6 (%29) | 8 (%38, Y501 is here) | 0 | 0 | 0 | 0 |
| 4 | 5 (%24, K484 is here) | 3 (%14) | 8 (%38, Y501 is here) | 8 (%38, Y501 is here) | 0 | 0 | 0 |
| 5 | 3 (%14, K484 is here) | 5 (%24) | 2 (%10) | 3 (%14, K417 is here) | 8 (%38, Y501 is here) | 0 | 0 |
| 6 | 3 (%14, K484 is here) | 5 (%24) | 2 (%10) | 3 (%14, K417 is here) | 4 (%19) | 4 (%19, Y501 is here) | 0 |
| 7 | 3 (%14, K484 is here) | 5 (%24) | 2 (%10) | 2 (%10) | 4 (%19) | 4 (%19, Y501 is here) | 1 (%5, K417 is here) |

As can be seen in the table, the variants we identified as dangerous were separated into 6 clusters. Therefore, we chose this number of clusters as optimal. From here, we investigated which variants would fall into the same cluster with the N501, K417 and E484K variants that we know as dangerous.

## 3. Results

Within the scope of the project, we obtained the mutation-induced S-protein stability and ACE2 binding change experimental data from Starr et al. This dataset includes 4221 S-protein mutations. Using our self-written python code, we extracted 586 experimental mutation data within this dataset that allowed or did not affect the S-protein binding to ACE2 better. We then encoded these mutations in terms of also Hydropathy, Polarity, Volume, Molecular weight, ring number in amino acid structure, oxygen number, hydrogen number and double bond number changes and correlated them with our experimental data. Our dataset and all pyton codes created in this way can be accessed from the link https://github.com/BiyoinformatikProje/Prediction-of-SARS-CoV-2-Mutations-that-Increase-Contagiousness . This set, which contains 586 mutations and includes N501Y and E484K data, was introduced to two different machine learning grouping algorithms. Next, for different group (cluster) numbers, it was investigated which mutations would fall into the same cluster with the N501Y and E484K mutations. Our goal here was to remove dangerous SARS-CoV-2 mutations that could cause changes similar to the N501Y and E484K mutations. For this purpose, we used K-means (k-means) clustering and EM (Expectation Maximization) algorithms, which give the most consistent results among all the clustering algorithms we have tried, through the Weka program, which is widely used for machine learning. The files of all analysis results of the K-mean and Expectation Maximization Algorithms within the scope of the project are given in the ANNEX-1.

### 3.1. K-mean clustering algorithm analysis results

Using the K-mean clustering algorithm, the whole data set was divided into different clusters with the number of clusters in the range of 2-12 (Table 2).

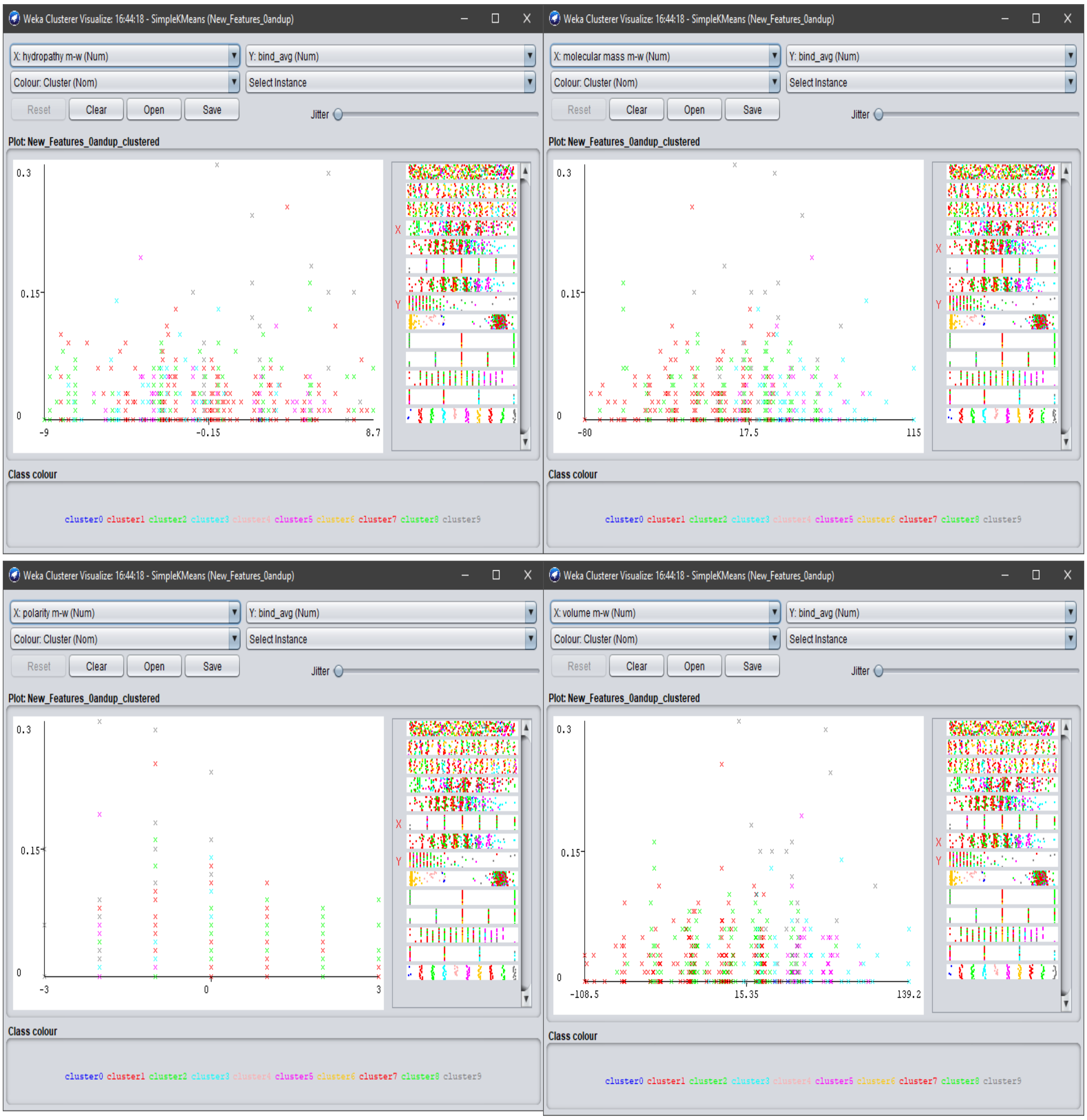
**Table 2.** K-means clustering algorithm results

| Number of Cluster | Cluster number where the E484K mutation is found | The cluster number where the N501Y mutation is located | The number of elements in the set with the E484K mutation | Number of elements in the set with the N501Y mutation |
|---|---|---|---|---|
| **2** | 1 | 1 | 445 | 445 |
| **3** | 1 | 1 | 299 | 299 |
| **4** | 3 | 3 | 164 | 164 |
| **5** | 3 | 3 | 164 | 164 |
| **6** | 5 | 3 | 118 | 66 |
| **7** | 5 | 3 | 119 | 65 |
| **8** | 1 | 1 | 57 | 57 |
| **9** | 8 | 8 | 56 | 56 |
| **10** | **9** | **9** | **30** | **30** |
| **11** | 9 | 9 | 30 | 30 |
| **12** | 9 | 9 | 29 | 29 |

Considering the clustering statistics given in Table 3, it is seen that the number of the cluster harboring N501Y and E484K mutations has not changed since K = 10. Considering this, it was decided to divide our data set into 10 separate clusters with the K-mean algorithm. In addition, it was observed that N501Y and E484K mutations fell to the same cluster for K = 2,3,4,5,8,9,1011,12 values in many clusters. This indicates the reliability of our clustering approach. As can be seen in Figures 8 and 9, the amino acid properties we use differently explain the binding of the S-protein to ACE2 and its stability. This shows the necessity of using each feature. When we look at the 9th cluster with 30 elements, the first five mutations with the highest stability were determined as follows:

**N501M, Q414A, Q498H, N460K, N501W (stability ranked from most to least)**

All of the mutations listed in this list have been identified as the five most dangerous mutations proposed by K-clustering. The fact that position 501 is listed twice in these mutations indicates that one should be careful about mutations that may occur at this position.


Weka Clusterer Visualize: 16:44:18 - SimpleKMeans (New_Features_0andup)
X: hydropathy m-w (Num)
Y: bind_avg (Num)
Colour: Cluster (Nom)
Select Instance
Plot: New_Features_0andup_clustered

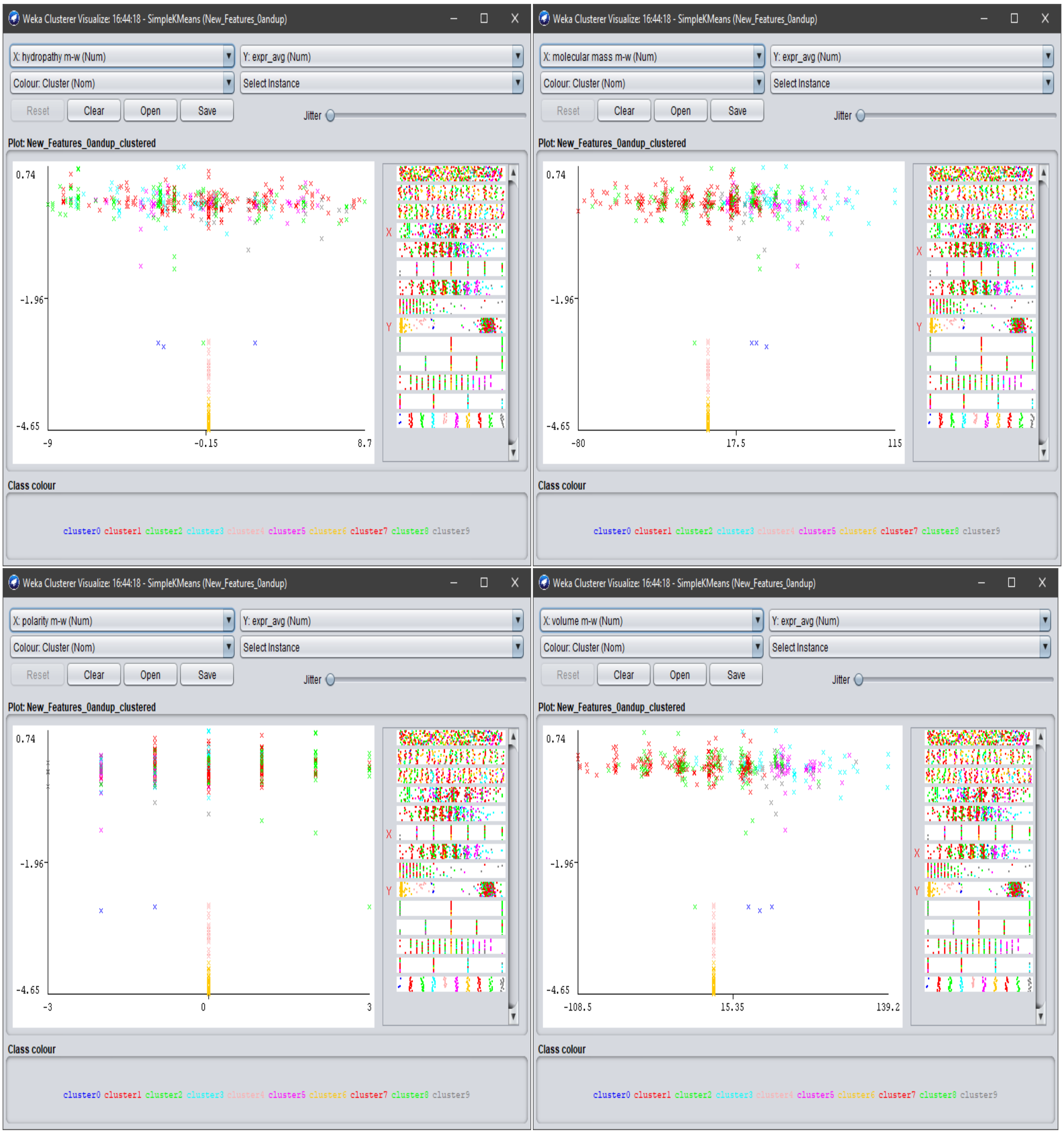
Weka Clusterer Visualize: 16:44:18 - SimpleKMeans (New_Features_0andup)
X: molecular mass m-w (Num)
Y: bind_avg (Num)
Colour: Cluster (Nom)
Select Instance
Plot: New_Features_0andup_clustered

Weka Clusterer Visualize: 16:44:18 - SimpleKMeans (New_Features_0andup)
X: polarity m-w (Num)
Y: bind_avg (Num)
Colour: Cluster (Nom)
Select Instance
Plot: New_Features_0andup_clustered

Weka Clusterer Visualize: 16:44:18 - SimpleKMeans (New_Features_0andup)
X: volume m-w (Num)
Y: bind_avg (Num)
Colour: Cluster (Nom)
Select Instance
Plot: New_Features_0andup_clustered

Reset
Clear
Open
Save
Jitter
0.3
0.15
0
-9
-0.15
8.7
-80
17.5
115
-3
0
3
-108.5
15.35
139.2
X
Y
Class colour
cluster0 cluster1 cluster2 cluster3 cluster4 cluster5 cluster6 cluster7 cluster8 cluster9

**Figure 9.** Results of the analysis performed to examine the binding effect of 4 different amino acid properties as a result of the K = 10 parameter using the K-mean clustering algorithm. Each plot has the bond strength on the y-axis (the higher the better) and on the x-axis there is left to right, hydropathy, molecular weight, polarity, and volume, respectively, from top to bottom. Data distributions are colored according to different groups.



Weka Clusterer Visualize: 16:44:18 - SimpleKMeans (New_Features_0andup)

X: hydropathy m-w (Num)
Y: expr_avg (Num)
Colour: Cluster (Nom)
Select Instance
Plot: New_Features_0andup_clustered

Weka Clusterer Visualize: 16:44:18 - SimpleKMeans (New_Features_0andup)

X: molecular mass m-w (Num)
Y: expr_avg (Num)
Colour: Cluster (Nom)
Select Instance
Plot: New_Features_0andup_clustered

Weka Clusterer Visualize: 16:44:18 - SimpleKMeans (New_Features_0andup)

X: polarity m-w (Num)
Y: expr_avg (Num)
Colour: Cluster (Nom)
Select Instance
Plot: New_Features_0andup_clustered

Weka Clusterer Visualize: 16:44:18 - SimpleKMeans (New_Features_0andup)

X: volume m-w (Num)
Y: expr_avg (Num)
Colour: Cluster (Nom)
Select Instance
Plot: New_Features_0andup_clustered

**Figure 10.** The results of the analysis performed to examine the effect of 4 different amino acid properties on the stability of the S-protein as a result of the K = 10 parameter with the K-mean clustering algorithm. Each plot has the bond strength on the y-axis (the higher the better), and on the x-axis there is left to right, hydropathy, molecular weight, polarity, and volume, respectively, from top to bottom. Data distributions are colored according to different groups.

### 3.2. Expectation Maximization clustering algorithm analysis results

Using the expectation maximization clustering algorithm, the whole data set was divided into different clusters with the number of clusters in the range of 2-12 (Table 3).

**Table 3.** Expectation maximization clustering algorithm results

| Number of Cluster | Cluster number where the E484K mutation is found | The cluster number where the N501Y mutation is located | The number of elements in the set with the E484K mutation | Number of elements in the set with the N501Y mutation |
|---|---|---|---|---|
| **2** | 1 | 1 | 449 | 449 |
| **3** | 0 | 0 | 342 | 342 |
| **4** | 0 | 0 | 175 | 175 |
| **5** | 0 | 0 | 63 | 63 |
| **6** | 2 | 2 | 57 | 57 |
| **7** | 2 | 2 | 61 | 61 |
| **8** | 4 | 4 | 56 | 56 |
| **9** | 4 | 4 | 51 | 51 |
| **10** | **8** | **8** | **46** | **46** |
| **11** | 3 | 3 | 49 | 49 |
| **12** | 3 | 3 | 49 | 49 |

Considering the clustering statistics given in Table 3, it is seen that the number of clusters containing N501Y and E484K mutations for this algorithm has increased since the value of K = 10. Considering this, it was decided that dividing our data set into 10 separate clusters with the EM algorithm is the most appropriate parametric case. In addition, it was observed that N501Y and E484K mutations fell into the same cluster in groupings. As can be seen in Figures 10 and 11, the amino acid properties we use differently explain the binding of the S-protein to ACE2 and its stability. When we look at the 8th cluster with 46 elements calculated for K = 10, the first five mutations with the highest stability were determined as follows:

**N501M, N354K, Q498H, N460K, N501W (stability ranked from most to least)**

Similar to the K-mean algorithm, in this algorithm, the same mutations at position 501 were selected in the list of the five most dangerous mutations.

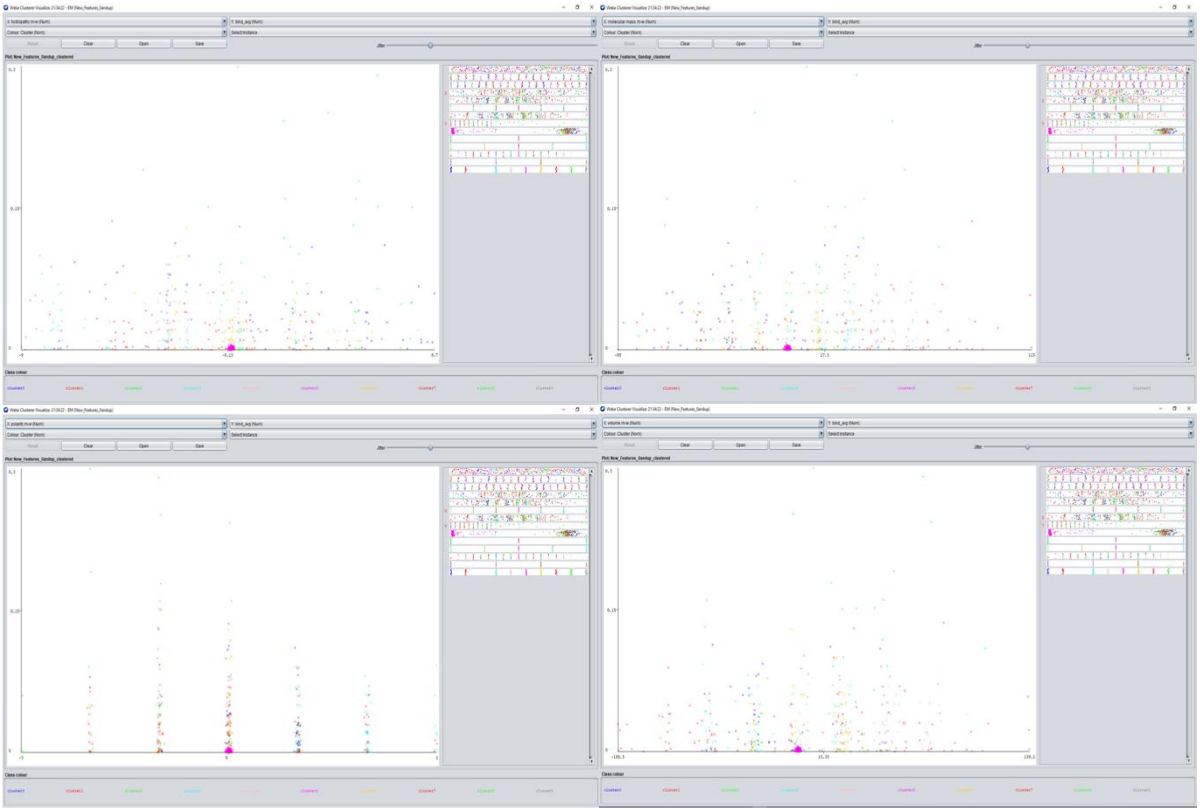


**Figure 11.** Analysis results for the purpose of examining the binding effect of the S-protein to ACE2 of 4 different amino acid properties as a result of the K = 10 parameter with the Expectation Maximization clustering algorithm. Each plot has the bond strength on the y-axis (the higher the better), and on the x-axis there is left to right, hydropathy, molecular weight, polarity, and volume, respectively, from top to bottom. Data distributions are colored according to different groups.

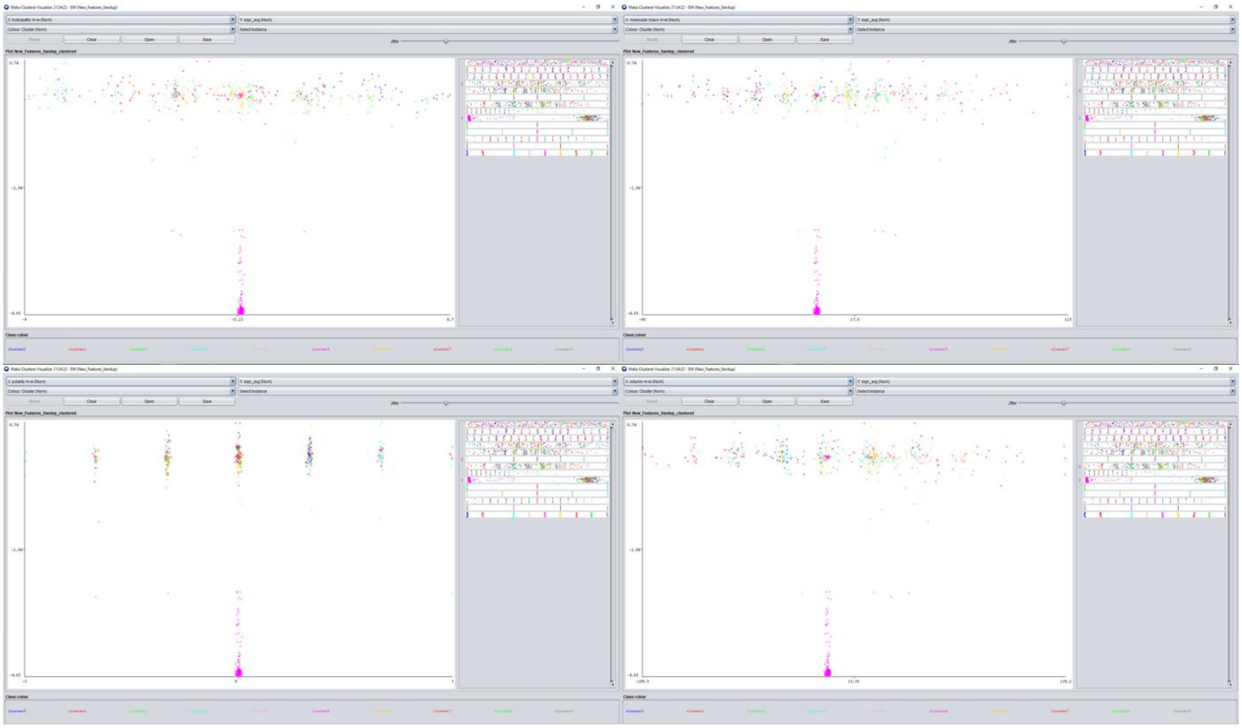


**Figure 12.** The results of the analysis performed to examine the effect of 4 different amino acid properties on the stability of the S-protein as a result of the K = 10 parameter with the Expectation Maximization clustering

algorithm. Each plot has the bond strength on the y-axis (the higher the better) and on the x-axis are left to right from top to bottom, hydropathy, molecular weight, polarity and volume. Data distributions are colored according to different groups.

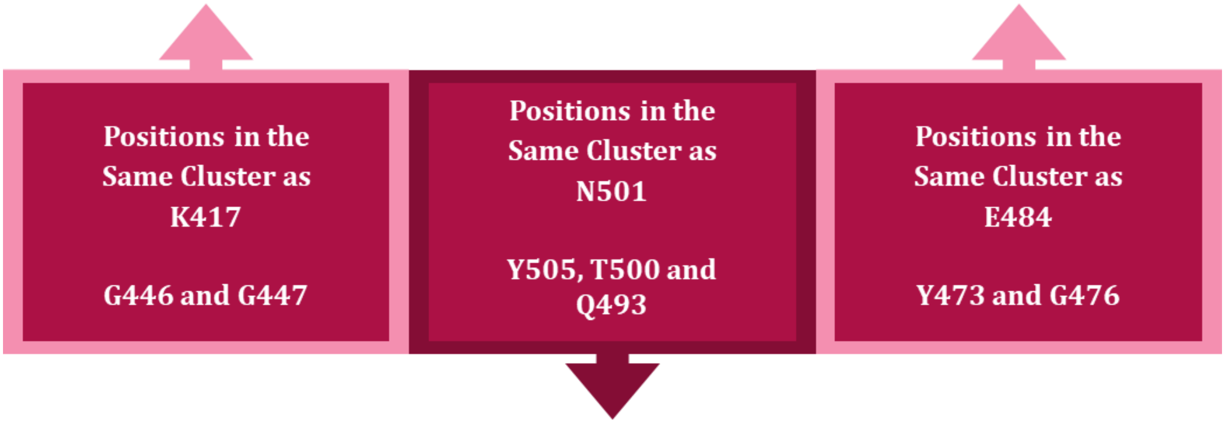
### 3.3. Common analysis results of EM and K-mean clustering algorithms

As a result of the analysis of the clusters obtained from the above two algorithms, it is seen that there are common mutations in the list of the five most dangerous mutations. With the evaluation of these common mutations, we consensusly identified six mutations as those that are likely to have dangerous effects on the spread of the coronavirus:

**N501M, Q414A, N354K, Q498H, N460K, N501W**

Special attention should be paid to the 501st position mutation seen in one of the currently widespread variants, we particularly suggest that position 501 repeats in this list.

### 3.4. Positions Where Dangerous Sars-Cov-2 Variants Can Be Seen



Positions in the Same Cluster as K417

G446 and G447

Positions in the Same Cluster as N501

Y505, T500 and Q493

Positions in the Same Cluster as E484

Y473 and G476

At the end of the analyzes performed on the multi-interaction data we created with the ACE2 AND RBD interaction data during the project development phase, the positions where dangerous variants can be seen were determined as G446, G447, Y505, T500, Q493, Y473 and G476.

## 4. Discussion

The global epidemic caused by SARS-CoV-2, a new type of human coronavirus, is a concern for all humanity. Considering the COVID-19 infection pandemic caused by the SARS CoV-2 virus, its scale and rapid spread, it appears to have become a major public health problem. Considering the worldwide dimension of the pandemic we are experiencing, the processes of working together between experts in different fields have been maximized in the race against time in the fight against coronavirus. In addition, it has come to the fore to follow approaches where different disciplines are combined and common sense is prioritized in order to find effective solutions urgently against the common threat. In this context, the research results to be obtained in the field of computational biology and machine learning will provide preliminary information to detailed laboratory studies, help in planning experiments, interpreting analyzes, and taking various treatment and health measures. For this purpose, within the scope of the project, SARS-CoV-2 mutations, which are commonly observed in

England, South Africa and Brazil and may cause a change similar to the N501Y and E484K mutations that cause concern all over the world, and which may pose a danger due to their high contagiousness. It was intended to be predicted.

✓ Experimental data for mutation-induced S-protein stability and binding change to ACE2 were obtained from the literature, and using the python code we wrote, 586 experimental mutation data that provided better binding of the S-protein to ACE2 or did not affect its binding were extracted. Later, the amino acid properties effective in these mutations were coded in terms of Hydropathy, Polarity, Volume, Molecular Weight, Number of Rings in Amino Acid Structure, Oxygen Number, Hydrogen Number and Double Bond Number Change and correlated with experimental data.

✓ This set, which contains 586 mutations and N501Y and E484K data, was introduced to two different machine learning grouping algorithms. Next, for different group (cluster) numbers, it was investigated which mutations would fall into the same cluster with the N501Y and E484K mutations. For this purpose, K-means (k-means) clustering and EM (Expectation Maximization) algorithms, which give the most consistent results among all the clustering algorithms we have tried, were used through the Weka program, which is widely used for machine learning. As a result of the analysis made in this context;

✓ When the 9 cluster with 30 elements calculated for k = 10 as a result of the analysis made with the K-mean clustering algorithm, the first five mutations with the highest stability were determined as N501M, Q414A, Q498H, N460K, N501W (ranked from the highest to the lowest). These mutations were identified as the five most dangerous mutations suggested by K-mean clustering. The fact that position 501 is listed twice in these mutations indicates that one should be careful about mutations that may occur at this position.

✓ As a result of the analysis performed with the expectation maximization clustering algorithm, when looking at the 8th cluster with 46 elements calculated for k = 10, the first five mutations with the highest stability were determined as N501M, N354K, Q498H, N460K, N501W (ranked from the highest to the lowest). Again in this algorithm, the same mutations belonging to position 501 were selected in the list of the five most dangerous mutations.

✓ It is seen that there are common mutations in the list of the five most dangerous mutations as a result of the analysis of the clusters obtained from both algorithms. Consensus with the evaluation of these common mutations, six mutations N501M, Q414A, N354K, Q498H, N460K, N501W were identified as highly likely to have dangerous effects on the spread of the coronavirus.

✓ Special attention should be paid to the 501st position mutation seen in one of the currently common variants, we particularly suggest that position 501 repeats in this list.

✓ At the end of the analyzes performed on the multi-interaction data we created with the ACE2 AND RBD interaction data during the project development phase, the positions where dangerous variants can be seen were determined as G446, G447, Y505, T500, Q493, Y473 and G476.

The scientific world is experiencing a dynamic pandemic process in which all kinds of studies to determine and predict the transmission rate and method of the virus will both provide serious benefits and be tested. Developing the possibility of accelerating the transmission of SARS-CoV-2 mutations in this area, especially with prediction algorithms, provides strong evidence that more rich and productive solutions will be obtained in the fight against the virus. Therefore, such studies will be a guide regarding the spread and progress of the virus.

## 5. Suggestions

Evaluating the impact of disease-related mutations has been of considerable interest at different levels. Computational methods for predicting the effect of mutations are an alternative to experimental techniques because they take less time and do not require biochemical work to prepare samples. For this, studies were carried out to estimate the contagiousness of future mutation types based on the information obtained from known SARS-CoV-2 mutations in the project. In the future studies of the project;

- ✓ In order to increase the number of factors examined, new amino acid properties can be added and their effects in algorithms can be examined.

- ✓ New associations can be made and interpreted by using different algorithms for changes in mutations based on amino acid properties.

- ✓ Mutation prediction algorithm can be developed.

- ✓ The efficiency of suggested mutations can be tested with advanced molecular simulation programs.

- ✓ As a result of the analysis carried out in the project, if mutations that fall into the same cluster with the N501Y and E484K mutations and are identified as the six most dangerous mutations, great sensitivity should be shown to prevent the spread of these mutations and necessary health measures should be taken.

## 6. References

1. Alpaydin, E. (2004). Introduction to machine learning, The MIT Press, ISBN 0-262-01211-1. (https://www.cmpe.boun.edu.tr/~ethem/i2ml/)
2. Amengual-Rigo, P., Fernández-Recio, J., & Guallar, V. (2020). UEP: an open-source and fast classifier for predicting the impact of mutations in protein–protein complexes. Bioinformatics.
3. Bhaskar L, Roshan B, Nasri H. The fuzzy connection between SARS-CoV-2 infection and loss of renal function. Am J Nephrol 2020;51:572e3.
4. Cucinotta D, Vanelli M. WHO Declares COVID-19 a pandemic. Acta Biomed : Atenei Parmensis 2020;91:157e60.
5. Gu, H., Chen, Q., Yang, G., He, L., Fan, H., Deng, Y. Q., ... & Zhou, Y. (2020). Adaptation of SARS-CoV-2 in BALB/c mice for testing vaccine efficacy. Science, 369(6511), 1603-1607.
6. Hoffmann, M., Kleine-Weber, H., & Pöhlmann, S. (2020). A multibasic cleavage site in the spike protein of SARS-CoV-2 is essential for infection of human lung cells. Molecular cell, 78(4), 779-784.
7. https://cran.r-project.org/web/packages/dendextend/vignettes/Cluster_Analysis.html
8. https://www.analyticsvidhya.com/blog/2020/10/the-mathematics-behind-svm/
9. https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning
10. https://www.npr.org/sections/goatsandsoda/2021/02/02/961668700/whats-going-on-with-all-these-coronavirus-variants-an-illustrated-guide
11. https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_clustering_algorithms_hierarchical.htm
12. https://www.tutorialspoint.com/weka/index.htm
13. https://www.who.int/emergencies/diseases/novel-coronavirus-2019
14. Jia, Y., Shen, G., Zhang, Y., Huang, K. S., Ho, H. Y., Hor, W. S., ... & Wang, W. L. (2020). Analysis of the mutation dynamics of SARS-CoV-2 reveals the spread history and emergence of RBD mutant with lower ACE2 binding affinity. BioRxiv.
15. Jiang F, Deng L, Zhang L, Cai Y, Cheung CW, Xia Z. Review of the clinical characteristics of coronavirus disease 2019 (COVID-19). J Gen Intern Med 2020;35:1545e9.
16. Kannan, S., ALI, P. S. S., Sheeza, A., & Hemalatha, K. (2020). COVID-19 (Novel Coronavirus 2019)recent trends. European Review for Medical and Pharmacolog ical Sciences, 24, 2006-2011.
17. Korber, B., Fischer, W. M., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W., ... & Montefiori, D. C. (2020). Tracking changes in SARS-CoV-2 Spike: evidence that D614G increases infectivity of the COVID-19 virus. Cell, 182(4), 812-827.
18. Lehninger, A. L., Nelson, D. L., & Cox, M. M. (2005). Lehninger principles of biochemistry. Macmillan.
19. Li X, Wang L, Yan S, Yang F, Xiang L, Zhu J, et al. Clinical characteristics of 25 death cases with COVID-19: a retrospective review of medical records in a single medical center, Wuhan, China. Int J Infect Dis 2020;94:128e32.
20. Li, G., Pahari, S., Murthy, A. K., Liang, S., Fragoza, R., Yu, H., & Alexov, E. (2020). SAAMBE-SEQ: a sequence-based method for predicting mutation effect on protein–protein binding affinity. Bioinformatics.
21. Luan, B., Wang, H., & Huynh, T. (2021). Molecular Mechanism of the N501Y Mutation

for Enhanced Binding between SARS-CoV-2's Spike Protein and Human ACE2 Receptor. bioRxiv, 2021-01.

**22.** Mahase E. Covid-19: death rate is 0.66% and increases with age, study estimates. BMJ 2020;369:m1327.

**23.** Nelson, G., Buzko, O., Spilman, P. R., Niazi, K., Rabizadeh, S., & Soon-Shiong, P. R. (2021). Molecular dynamic simulation reveals E484K mutation enhances spike RBD-ACE2 affinity and the combination of E484K, K417N and N501Y mutations (501Y. V2 variant) induces conformational change greater than N501Y mutant alone, potentially resulting in an escape mutant. bioRxiv.

**24.** Pachetti, M., Marini, B., Benedetti, F., Giudici, F., Mauro, E., Storici, P., Masciovecchio, C., Angeletti, S., Ciccozzi, M., Gallo, R., Zella, D., Ippodrino, R., 2020. Emerging SARS CoV-2 mutation hot spots ınclude a novel RNAdependent-RNA polymerase variant. Journal of Translational Medicine, 18:1–9.

**25.** Phan, T., 2020. Genetic diversity and evolution of SARS-CoV-2. Infection, Genetics and Evolution, 81:1-3.

**26.** Rajgor DD, Lee MH, Archuleta S, Bagdasarian N, Quek SC. The many estimates of the COVID-19 case fatality rate. Lancet Infect Dis 2020;20:776e7.

**27.** Rodrigues, J. P., Barrera-Vilarmau, S., Mc Teixeira, J., Sorokina, M., Seckel, E., Kastritis, P. L., & Levitt, M. (2020). Insights on cross-species transmission of SARS-CoV-2 from structural modeling. PLoS computational biology, 16(12), e1008449.

**28.** Seah, I., Su, X., & Lingam, G. (2020). Revisiting the dangers of the coronavirus in the ophthalmology practice.

**29.** Shang, J., Ye, G., Shi, K., Wan, Y., Luo, C., Aihara, H., ... & Li, F. (2020). Structural basis of receptor recognition by SARS-CoV-2. Nature, 581(7807), 221-224.

**30.** Stamp, M. (2018). A survey of machine learning algorithms and their application in information security. In Guide to Vulnerability Analysis for Computer Networks and Systems (pp. 33-55). Springer, Cham.

**31.** Starr, T. N., Greaney, A. J., Hilton, S. K., Ellis, D., Crawford, K. H., Dingens, A. S., ... & Bloom, J. D. (2020). Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. Cell, 182(5), 1295-1310.

**32.** Sun, T., Zhou, B., Lai, L., & Pei, J. (2017). Sequence-based prediction of protein protein interaction using a deep-learning algorithm. BMC bioinformatics, 18(1), 1-8.

**33.** Valizadeh R, Baradaran A, Mirzazadeh A, Bhaskar LV. Coronavirus-nephropathy; renal involvement in COVID-19. J Ren Inj Prev 2020;9:18.

**34.** Van Dorp, L., Acman, M., Richard, D., Shaw, L., Ford, C., Ormond, L., Owen, C., Pang, J., Tan, C., Boshier, F., Ortiz, A., Balloux, F., 2020. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. Infection, Genetics and Evolution, 83:104351

**35.** Verma, H. K., Merchant, N., Verma, M. K., Kuru, C. İ., Singh, A. N., Ulucan, F., ... & Bhaskar, L. V. K. S. (2020). Current updates on the European and WHO registered clinical trials of coronavirus disease 2019 (COVID-19). Biomedical journal.

**36.** Walls, A. C., Park, Y. J., Tortorici, M. A., Wall, A., McGuire, A. T., & Veesler, D. (2020). Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. Cell, 181(2), 281-292.

**37.** World Health Organization. COVID-19 weekly epidemiological update.

## 【評語】080012

The aim of this project is to predict which SARS-CoV-2 mutations will increase the viral infectiousness in humans by using the machine learning WEKA program. The authors investigated the interactions between amino acid change properties of viral spike protein and its human target protein ACE2 receptor. In this study, bioinformatic approaches predicted that the following mutations at RBD, including N354, Q414, G446, G447, N460, Y473, G476, Q493, Q498, T500, N501 and Y505, may increase contagiousness. However, many predicted mutational sites were not observed in the Omicron variant that recently causes pandemic outbreak.

Suggestions：

1. It is recommended to define all abbreviations which have been introduced first time they occur in the text. For example, ACE2 and RBD appear late on page 2.
2. Does the alpha variant of SARS-CoV-2 have spike N501Y and E484K mutations？ Omicron variant has more than 30 amino-acid changes on spike protein, and half of these 30 changes are located in the RBD. Are these Omicron changes successfully predicted by your current study？