Смысл всего данного проекта в том, чтобы, получив раздробленные цепочки РНК определить каким вирусням (в моем случае) они принадлежат из тех, кто у меня есть базе, на которой я и производил обучение.

Для начала я получаю fasta файлы, в которых хранятся цепочки РНК в моем случает вирусов ковид.

**I) Аппликацию create\_dataset.py я использую для того, чтобы:**

1. Получить ген с цепочек – это элементы [21563:25385];
2. Разбить цепочку гена на k-mer длиной 21 и запись этих данных в датафрэйм со структурой:
   1. ref – имя вируса;
   2. ref\_index – индекс вируса;
   3. pos – позиция k-mer в цепочке гена;
   4. val – k-mer;
   5. val\_num – индекс k-mer;

Имя датафрэйма data\_total.

1. После я записываю данные с data\_total[‘val’] в другой датасет и нахожу дубликаты по значению k-mer и удаляю эти строки, чтобы в итоге осталась только одна строка с каждым k-mer и задаю им уникальные (в моем случае порядковые) индексы и записываю эти данные в файл keys.csv. Это ассоциативный файл чтобы в наборе, который я получу для тестирования аппликации я знал какому k-mer какой номер присвоить, под которым он проходил в тренировке;
2. Cам датафрэйм data\_total я записываю в файл dataset.csv.

**II) Аппликацию create\_samples.py я использую для того, чтобы**:

1. Создаю датафрэйм data\_total со структурой:
   1. pos – позиция k-mer;
   2. val – значение k-mer;
   3. val\_num – числовое значение k-mer, которое я беру из файла keys.csv.
2. С файла R.csv я считываю целую цепочку;
3. Делю цепочку на k-mer с длиной 21;
4. Используя get\_kmer\_from\_gene\_str я записываю данные в data\_total;
5. Используя keys.csv я добавляю val\_num(индекс такого же k-mer, участвовавшего в трейнинге);
6. Записываю data\_total в samples.csv.

**III) Аппликацию main.py я использую для того, чтобы сделать саму классификацию**:

1. Данные из пункта I из файла dataset.csv я загружаю и использую для обучения;
2. Данные из пункта II из файла samples.csv я загружаю для данных, которые я буду классифицировать, т.е. для предсказания;
3. Результат прогноза я записываю в файл samples\_pred.csv, добавив столбец y\_pred, под которым я записал индекс предсказанного(классифицированного) вируса;
4. После этого я делаю предикт с данными из dataset.csv, предварительно разделив их на трейнинг и тест в соотношении 80% на 20% и вывожу accuracy\_score, который у меня около 60% и confusion\_matrix.