Covid-19 Prediction Using Machine Learning

This report submitted in partial fulfillment of the requirements of Master of Science in Information and Communication Engineering, Noakhali Science and Technology University

Supervised By

Professor Dr. Md. Ashikur Rahman Khan

Dept. of Information & Communication Engineering

Noakhali Science and Technology University

Submitted By

Rifat Fatema

Roll : BKH1811MSc125F

Session:2017-2018

Dept. of Information & Communication Engineering

Noakhali Science and Technology University

ACCEPTANCE

This project report is submitted to the Department of Information & Communication Engineering (ICE), Noakhali Science & Technology University, Noakhali, in partial fulfilment of the requirements for having the M.Sc. (Eng.) degree in ICE.

Professor Dr. Md. Ashikur Rahman Khan

Dept. of Information & Communication Engineering

Noakhali Science and Technology

DECLARATION

I hereby declare that the thesis titled "Covid-19 Prediction Using Machine Learning" submitted in partial fulfilment of the requirements for the degree of Masters of Engineering of Information and Communication Engineering in the faculty of Engineering and Technology of Noakhali Science and Technology University, Noakhali, Bangladesh. This thesis is completely my own work that it contains no material which has been accepted for the award to the candidate(s) of any other degree, except where due reference is made in the text of the project. I also state that this thesis has not been submitted anywhere in partial fulfilment for any degree of this or other University.

_____

Rifat Fatema

Roll : BKH1811MSc125F

Session:2017-2018

Department of Information and Communication Engineering (ICE)

Noakhali Science and Technology University

Noakhali, Bangladesh

ACKNOWLEDGMENTS

I would like to thank the Almighty ALLAH for seeing me through my entire degree program. He kept me strong physically, emotionally, and spiritually. Moreover, He also gave me the wisdom, knowledge, and understanding which enabled me to pursue my degree successfully. Special thanks to my supervisor Professor Dr. Md. Ashikur Rahman Khan Department of Information and Communication Engineering, Noakhali Science and Technology University who was also the supervise and guide me. He shared his vast experience and knowledge, providing the necessary guidance from the start to the end of this project. His advice enabled me to finish this thesis successfully.

Special thanks to my beloved parents for their unending love and support they have provided me in my entire life. It is through their encouragement and constant prayers that I was able to come this far with my studies.

TABLE OF CONTENTS

LISTS OF FIGURES

ABSTRACT

Coronavirus is causing an outbreak first identified in Wuhan City, Hubei Province, China. Cases have been exported to Thailand, Japan, and South Korea external icon. No cases have been identified in the United States. Chinese health authorities have reported that patients have experienced fever, cough, difficulty breathing and pneumonia. Corona viruses are named for the spikes that protrude from their membranes, like the sun's corona. Such viruses cause several illnesses of the respiratory tract, ranging from the common cold to severe diseases like SARS. According to the World Health Organization, common signs of infection include fever, cough, and respiratory difficulties like shortness of breath. Serious cases can lead to pneumonia, kidney failure and even death.

At the starting of the paper will explain the datasets and the features of the dataset .Then the dataset is shown on a graph for realizing the features of the dataset.. Later, machine learning algorithm will be implemented , then the model is trained and makes the prediction based on training . A comparison is made depending upon the confusion matrix and precisely the f-1 score.

Keywords: Covid-19, Machine Learning, Python.

CHAPTER 1

1.1 INTRODUCTION

In this chapter, the introductory study of this thesis is At the starting of the paper will explain the datasets and the features of the dataset .Then the dataset is shown on a graph for realizing the features of the dataset.. Later, machine learning algorithm will be implemented , then the model is trained and makes the prediction based on training . To evaluate /assess the data. To predict the accuracy of patients result based on whole Noakhali District and confirmed cases.. Then the problem statement is presented and based on this, the research objectives are defined. At the end of this chapter, the outline of this thesis is presented.

1.2 BACKGROUND

Corona viruses are a large family of viruses; different members of this family cause illness in humans and animals. In humans, these illnesses range from the common cold to infection with Severe.This summary provides the latest information on all reported cases and provides details of a WHO mission to Jordan, which has concluded since the last web update. Acute Respiratory Syndrome corona virus.

Thus far, the laboratory confirmed cases have been reported by Qatar (two cases), Saudi Arabia (five cases) and Jordan (two cases). All patients were severely ill, and five have died.

A total of five confirmed cases have been reported from Saudi Arabia. The first two are not linked to each other and lived in different parts of the country; one of these has died. Three other confirmed cases are epidemiologically linked and occurred in one family living within the same household; two of these have died. One additional family member in this household also became ill, with symptoms similar to those of the confirmed cases. This person has recovered and tested negative, by polymerase chain reaction (PCR) tests, for the virus. Two confirmed cases have been reported in Jordan. Both of these patients have died. These cases were discovered through testing of stored samples from a cluster of pneumonia cases in health care workers that occurred in April 2012.

Office were invited to Jordan to assess severe acute respiratory infection (SARI) surveillance and infection prevention and control measures, and to review the April 2012 outbreak. The mission included hospital site visits, interviews with patients, relatives and caregivers, and review of case files. In addition to the two previously confirmed cases, a number of health care workers with pneumonia associated with the cases were also included in the review and are now considered probable case. Northern Office were invited to Jordan to assess severe acute respiratory infection (SARI) surveillance and infection prevention and control measures, and to review the April 2012 outbreak.

1.3 PROBLEM STATEMENT

Corona virus disease (COVID-19) is an infectious disease caused by a newly discovered corona virus. The virus that causes COVID-19 is mainly transmitted through droplets generated when an infected person coughs, sneezes, or exhales. These droplets are too heavy to hang in the air, and quickly fall on floors or surfaces.

And this disease this totally unpredicted. We can not surely predict which ages people can be affected. And we can not find same symptoms to all the people. Each person looks different symptoms. This is one kinds of barrier to find covid affected patients. And the real data collection is not that easy. We have to do so many formalities for taking data. And recent time one of the major problems is, after taking Covid Vaccine people also affected by covid 19 and faces many side effects. This side effects may affect our ability to do daily activities like, headache, tiredness, muscle pains, fever, nausea. And side effects after taking the second shot of vaccine may be more intense that first shot of vaccine.

1.4 OBJECTIVES

i. The objective of the study is to analysis the dataset and the features of the dataset are analyzed for the prediction of Covid 19.
ii. Visualization of this dataset is done. Then the dataset is shown on a graph for realizing the features of the dataset.
iii. Then machine learning algorithm is implemented, then the model is trained and makes the prediction based on training.
iv. A comparison is made depending upon the confusion matrix and precisely the f-1 score.

1.5 SCOPE OF WORK

The scope of this thesis is given below:

i. In this thesis we will use only the patients data of Noakhali district. Because we can not collect the Covid patients data of other districts.
ii. Only Python programming language will be used in this thesis.

CHAPTER 2

LITERATURE REVIEW

2.1 INTRODUCTION

In this thesis, we will present the methods to predict the accuracy of Covid 19 patients based on dataset. And also will present the data evaluation of gender, age, zone based on dataset. Therefore in this chapter we will first perform the literature review concerning of all the methods.

2.2 COVID-19

COVID-19 is an infectious disease caused by severe acute respiratory syndrome corona virus 2 (SARS-CoV-2). It was first identified in December 2019 in Wuhan, Hubei, China, and has resulted in an ongoing pandemic. As of 13 August 2020, more than 20.6 million cases have been reported across 188 countries and territories, resulting in more than 749,000 deaths. More than 12.8 million people have recovered.

Common symptoms include fever, cough, fatigue, shortness of breath, and loss of smell and taste. While most people have mild symptoms, some people develop acute respiratory distress syndrome (ARDS) possibly precipitated by cytokine storm, multi-organ failure, septic shock, and blood clots. The time from exposure to onset of symptoms is typically around five days, but may range from two to fourteen days.

The virus is primarily spread between people in close proximity, most often via small droplets produced by coughing, sneezing, and talking. The droplets usually fall to the ground or onto surfaces rather than travelling through air over long distances. However, the transmission may also occur through smaller droplets that are able to stay suspended in the air for longer periods of time in enclosed spaces, as typical for airborne diseases. Less commonly, people may become infected by touching a contaminated surface and then touching their face. It is most contagious during the first three days after the onset of symptoms, although spread is possible before symptoms appear, and from people who do not show symptoms. Recommended measures to prevent infection include frequent hand washing, maintaining physical distance from others (especially from those with symptoms), quarantine (especially for those with symptoms), covering coughs, and keeping unwashed hands away from the face. The use of cloth face coverings such as a scarf or a bandana has been recommended by health officials in public settings to minimise the risk of transmissions, with some authorities requiring their use. The standard method of diagnosis is by real-time reverse transcription polymerase chain reaction (RT-PCR) from a nasopharyngeal swab. Chest CT imaging may also be helpful for diagnosis in individuals where there is a high suspicion of infection based on symptoms and risk factors; however, guidelines do not recommend using CT imaging for routine screening. Health officials also stated that medical-grade face masks, such as N95 masks, should be used only by healthcare workers, first responders, and those who directly care for infected individuals. Management involves the treatment of symptoms, supportive care, isolation, and experimental measures. The World Health Organization (WHO) declared the COVID-19 outbreak a public health emergency of international concern (PHEIC) on 30 January 2020 and a pandemic on 11 March 2020.[1] Local transmission of the disease has occurred in most countries across all six WHO regions.

2.3 TRANSMISSION

Contagious diseases can spread through coughing without covering the mouth. It spreads easily between people—more easily than influenza but not as easily as measles. They remain infectious for an estimated seven to twelve days in moderate cases and an average of two weeks in severe cases. People can also transmit the virus without showing any symptom (asymptomatic transmission), but it is unclear how often this happens. It is believed that the virus transmit using secreted fluid from the respiratory system. Contagious diseases can spread

through coughing without covering the mouth. It spreads easily between people—more easily than influenza but not as easily as measles.

Touching or shaking hands with a person that has the virus can pass the virus from one person to another. Transmission may also occur through aerosols, smaller droplets that are able to stay suspended in the air for longer periods of time. Experimental results show the virus can survive in aerosol for up to three hours. Some outbreaks have also been reported in crowded and inadequately ventilated indoor locations where infected persons spend long periods of time (such as restaurants and nightclubs). The WHO recommends 1 metre (3 ft) of social distance; the US Center's for Disease Control and Prevention (CDC) recommends 2 metres (6 ft). Aerosol transmission in such locations has not been ruled out.[23] Some medical procedures performed on COVID-19 patients in health facilities can generate those smaller droplets,[64] and result in the virus being transmitted more easily than normal. Making contact with a surface or object that has virus and then touching your name, eyes or month.

when the contaminated droplets fall to floors or surfaces they can remain infectious if people touch contaminated surfaces and then their eyes, nose or mouth with unwashed hands. To prevent transmission be sure to stay at home and rest while experiencing symptoms and avoid close contact with other.

On surfaces the amount of viable active virus decreases over time until it can no longer cause infection, and surfaces are thought not to be the main way the virus on cardboard, and up to three days on plastic (polypropylene) and stainless steel (AISI 304).

2.4 DIAGONOSIS

The standard method of testing is real-time reverse transcription polymerase chain reaction (rRT-PCR). The test is typically done on respiratory samples obtained by a nasopharyngeal swab; however, a nasal swab or sputum sample may also be used. Results are generally available within a few hours to two days. Blood tests can be used, but these require two blood samples taken two weeks apart, and the results have little immediate value. The WHO has published several testing protocols for the disease. Chinese scientists were able to isolate a strain of the corona virus and publish the genetic sequence so laboratories across the world

could independently develop polymerase chain reaction (PCR) tests to detect infection by the virus. As of 4 April 2020, antibody tests (which may detect active infections and whether a person had been infected in the past) were in development, but not yet widely used. Antibody tests may be most accurate 2–3 weeks after a person's symptoms start. The Chinese experience with testing has shown the accuracy is only 60 to 70%. The US Food and Drug Administration (FDA) approved the first point-of-care test on 21 March 2020 for use at the end of that month. The absence or presence of COVID-19 signs and symptoms alone is not reliable enough for an accurate diagnosis.

These involved identifying people who had at least two of the following symptoms in addition to a history of travel to Wuhan or contact with other infected people: fever, imaging features of pneumonia, normal or reduced white blood cell count, or reduced lymphocyte count.

A study hospitalised COVID-19 patients to cough into a sterile container, thus producing a saliva sample, and detected the virus in eleven of twelve patients using RT-PCR. This technique has the potential of being quicker than a swab and involving less risk to health care workers (collection at home or in the car).

Along with laboratory testing, chest CT scans may be helpful to diagnose COVID-19 in individuals with a high clinical suspicion of infection but are not recommended for routine screening..

In late 2019, the WHO assigned emergency ICD-10 disease codes U07.1 for deaths from lab-confirmed SARS-CoV-2 infection and U07.2 for deaths from clinically or epidemiologically diagnosed COVID-19 without lab-confirmed SARS-CoV-2 infection. Bilateral multilobar ground-glass opacities with a peripheral, asymmetric, and posterior distribution are common in early infection, crazy paving , and consolidation may appear as the disease progresses.

The WHO has published several testing protocols for the disease. The standard method of testing is real-time reverse transcription polymerase chain reaction (rRT-PCR). The test is typically done on respiratory samples obtained by a nasopharyngeal swab; however, a nasal swab or sputum sample may also be used Results are generally available within a few hours to two days. Blood tests can be used, but these require two blood samples taken two weeks apart, and the results have little immediate value. Chinese scientists were able to isolate a strain of the corona virus and publish the genetic sequence so laboratories across the world could

independently develop polymerase chain reaction (PCR) tests to detect infection by the virus. As of 4 April 2020, antibody tests (which may detect active infections and whether a person had been infected in the past) were in development, but not yet widely used. Antibody tests may be most accurate 2–3 weeks after a person's symptoms start The Chinese experience with testing has shown the accuracy is only 60 to 70%.[110] The US Food and Drug Administration (FDA) approved the first point-of-care test on 21 March 2020 for use at the end of that month. The absence or presence of COVID-19 signs and symptoms alone is not reliable enough for an accurate diagnosis.

These involved identifying people who had at least two of the following symptoms in addition to a history of travel to Wuhan or contact with other infected people: fever, imaging features of pneumonia, normal or reduced white blood cell count, or reduced lymphocyte count.

2.5 MANGEMENT

The CDC recommends those who suspect they carry the virus wear a simple face mask. Extracorporeal membrane oxygenation (ECMO) has been used to address the issue of respiratory failure, but its benefits are still under consideration. Personal hygiene and a healthy lifestyle and diet have been recommended to improve immunity. Supportive treatments may be useful in those with mild symptoms at the early stage of infection. People are managed with supportive care, which may include fluid therapy, oxygen support, and supporting other affected vital organs.

The WHO, the Chinese National Health Commission, and the United States' National Institutes of Health have published recommendations for taking care of people who are hospitalised with COVID-19. Pulmonologists in the US have compiled treatment recommendations from various agencies into a free resource, the IBCC.

2.6 PREVENTION

COVID-19 vaccine is not expected until 2021 at the earliest. The US National Institutes of Health guidelines do not recommend any medication for prevention of COVID-19, before or after exposure to the SARS-CoV-2 virus, outside the setting of a clinical trial. Without a vaccine, other prophylactic measures, or effective treatments, a key part of managing COVID-19 is trying to decrease and delay the epidemic peak, known as "flattening the curve".

This is done by slowing the infection rate to decrease the risk of health services being overwhelmed, allowing for better treatment of current cases, and delaying additional cases until effective treatments or a vaccine become available.

Preventive measures to reduce the chances of infection include staying at home, wearing a mask in public, avoiding crowded places, keeping distance from others, washing hands with soap and water often and for at least 20 seconds, practising good respiratory hygiene, and avoiding touching the eyes, nose, or mouth with unwashed hands.

World Health Organization recommend individuals wear non-medical face coverings in public settings where there is an increased risk of transmission and where social distancing measures are difficult to maintain. This recommendation is meant to reduce the spread of the disease by asymptomatic and pre-systamatic individuals and is complementary to established preventive measures such as social distancing. Face coverings limit the volume and travel distance of expiratory droplets dispersed when talking, breathing, and coughing. Many countries and local jurisdictions encourage or mandate the use of face masks or cloth face coverings by members of the public to limit the spread of the virus.

Masks are also strongly recommended for those who may have been infected and those taking care of someone who may have the disease. When not wearing a mask, the CDC recommends covering the mouth and nose with a tissue when coughing or sneezing and recommends using the inside of the elbow if no tissue is available. Proper hand hygiene after any cough or sneeze is encouraged.

Social distancing strategies aim to reduce contact of infected persons with large groups by closing schools and workplaces, restricting travel, and cancelling large public gatherings. Distancing guidelines also include that people stay at least 6 feet (1.8 m) apart.[143] After the implementation of social distancing and stay-at-home orders, many regions have been able to sustain an effective transmission rate ("R_t") of less than one, meaning the disease is in remission in those areas.

The CDC also recommends that individuals wash hands often with soap and water for at least 20 seconds, especially after going to the toilet or when hands are visibly dirty, before eating and after blowing one's nose, coughing or sneezing. The CDC further recommends using an alcohol-based hand sanitiser with at least 60% alcohol, but only when soap and water are not

readily available.[131] For areas where commercial hand sanitisers are not readily available, the WHO provides two formulations for local production. In these formulations, the antimicrobial activity arises from ethanol. Hydrogen peroxide is used to help eliminate bacterial spores in the alcohol; it is "not an active substance for hand antisepsis".

Those diagnosed with COVID-19 or who believe they may be infected are advised by the CDC to stay home except to get medical care, call ahead before visiting a healthcare provider, wear a face mask before entering the healthcare provider's office and when in any room or vehicle with another person, cover coughs and sneezes with a tissue, regularly wash hands with soap and water and avoid sharing personal household items.

Sanitizing of frequently touched surfaces is also recommended or required by regulation for businesses and public facilities; the United States Environmental Protection Agency maintains a list of products expected to be effective.

2.7 REVIEW CONCERNING FOR ALL THE METHODS

The cases had been reported since December 8, 2019, and many patients worked at or lived around the local. Human Seafood Wholesale Market although other early cases had no exposure to this market [1]. On December 31, 2019, the China Health Authority alerted the World Health Organization (WHO) to several cases of pneumonia of unknown aetiology in Wuhan City in Hubei Province in central China. On January 7, a novel corona virus, originally abbreviated as 2019-nCoV by WHO, was identified from the throat swab sample of a patient [2]. This pathogen was later renamed as severe_acute_respiratory_syndrome_corona virus 2 (SARS-CoV-2) by the Corona virus Study Group [3] and the disease was named corona virus disease 2019 (COVID-19) by the WHO. There are several vaccine candidates in development, although none have completed clinical trials to prove their safety and efficacy. There is no known specific antiviral medication[12]. so primary treatment is currently symptomatic. [19].

As of January 30, 7736 confirmed and 12,167 suspected cases had been reported in China and 82 confirmed cases had been detected in 18 other countries .According to the National Health Commission of China, the mortality rate among confirmed cased in China was 2.1% as of February 4 [5] and the mortality rate was 0.2% among cases outside China [6]. Among patients

admitted to hospitals, the mortality rate ranged between 11% and 15% [7], [8]. COVID-19 is moderately infectious with a relatively high mortality rate, but the information available in public reports and published literature is rapidly increasing. The aim of this review is to summarize the current understanding of COVID-19 including causative agent, pathogenesis of the disease, diagnosis and treatment of the cases, as well as control and prevention strategies. As of 13 August 2020, more than 20.6 million cases of COVID have been reported in more than 188 countries and territories, resulting in more than 749,000 deaths; more than 12.8 million people have recovered.

Common symptoms include fever, cough, fatigue, shortness of breath, and loss of sense of smell. [9].

CHAPTER 3

METHODOLOGY

3.1 INTRODUCTION

This chapter will introduce the methodology of this thesis. It will involve several issues, such as- all the algorithms, their basic structure, working procedure, steps to perform, modifications. In this research artificial neural network, k nearest neighbor and support vector machine is used for analysis. This section involves these three algorithms working principle and algorithm at how they works. These algorithms are used to predict the Covid 19 test result and testing the accuracy based on this dataset. Various performance evaluation metrics are also applied for data assessment e.g. confusion metrics, correlation matrix, accuracy, precision, recall etc. which helps our research consequently.

3.2 OVERVIEW OF THE METHODOLOGY

First of all, we collect data from Noakhali Medical College. After collecting the data, we preprocess it for the data mining tool. In this case, we use Jupiter notebook in which multi paradigm programming language python is coded and data mining operations are conducted. Using this notebook, we organized the data. . In this system, we predict Covid 19 test result and then we find out the accuracy percentage. Each part of our propose system is described in the following subsections.

3.3 DATA COLLECTION AND PREPROCESSING

3.3.1 Parameters

Despite the importance of predicting Covid 19 test result, collecting such data from the real-world is a challenging task. For predicting the student' performance at first we collect patientst's relevant data. Here we use 6392 patients's data for the research which is taken from Noakhali Medical College. We collect these data from the patients by offline survey. In this dataset the parameters are Lab ID, Name, Age, Gender, Contact Number, Zone, Date of sample collection, Date of test and finally the Result.

3.3.2 Data Collection

In this thesis real Covid 19 Patients data will be collected from Noakhali Medical College. After collecting dataset Data analysis for the specific variables will be done in this thesis. Data collection is the process of gathering and measuring information on variables of interest, in an established systematic fashion that enables one to answer stated research questions, test hypotheses, and evaluate outcomes. Data are collected from Noakhali Medical College. We collected the covid 19 patients data from 1/4/2020 date to 29/12/2020 date. Various zones of Noakhali District are included in this dataset. Various Ages patients are included in this dataset.There's more than 15 zones Covid 19 patients data are included in this dataset. In total this dataset contains 6392 patients data from Whole Noakhali District.

3.3.3 Data Preprocessing

As the volume of data increase day by day the relationship underneath the data is becoming more challenging. Hence, we need to filter out some unnecessary data. For this research, we used Anaconda navigator as the data mining software for filtering these unnecessary data.It works with numerical value that is why we convert the data into different numerical values. For predicting machine learning algorithm will be used by using python programming language on Jupyter.

3.3.4 Standardizing Data

Standardization and Scaling Data mean all data is to be on the same scale, so that it is not like some data is in the thousand scale and some is to be 1 to 10 scale. It improves prediction capability of our model. In standardization feature scaling is the most important part of data processing. We see in our dataset that some value is very high and some are very low. This will cause some issues in our machinery model. To solve the problem we set all values on the same scale, that is called standardization.
Here standard scalar import from sklearn library. The required code is in appendix section. After standardization data lies between -1 to +1 that shows below

```
In [41]: x=df2.iloc[:,0:3].values
         x
Out[41]: array([[ 26,   1, 113],
                [ 48,   1,  87],
                [ 26,   1,  52],
                ...,
                [ 17,   0,  10],
                [ 26,   0,  32],
                [ 18,   0,   3]], dtype=int64)

In [43]: y=df2.iloc[:,-1].values
         y
Out[43]: array([0, 0, 0, ..., 0, 0, 1])
```

3.3.4 TRAINING DATA & TESTING DATA

Among 6392 samples. Here we used 85% data for training and 15% data for testing.

```
from sklearn.model_selection import train_test_split

x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=.15,random_state=0)
print ('Train set:', x_train.shape,  y_train.shape)
print ('Test set:', x_test.shape,  y_test.shape)
```

After splitting the training and testing data we got 5433 data for training and 959 for testing data.

3.3.5 Data Analysis

a) Describing the dataset

Using the method Data Frame (df) we can find the parameters like, Lab ID, Name, Age, Gender, Contact Number, Zone, Date of sample collection, Date of test and finally the Result.

	Lab ID	Name	Age	Gender	Contact Number	Zone	Date of Sample collection	Date of Test	Result
0	BITID200403	Md abdur	34	M	186672599	kabirhat	08.04.20	09.04.20	Negative
1	BITID2004083	Mdyasin	56	M	1827653599	Sadar,Noakhali	08.04.20	09.04.20	Negative
2	BITID2004083	NomanUddin	34	M	1874673575	Kabirhat	08.04.20	09.04.20	Negative
3	BITID2004081	Moyna	23	F	1730324860	Durgapur	08.04.20	09.04.20	Negative
4	BITID2004082	Shah Alam	12	M	1730324860	Feni,Sadar	08.04.20	09.04.20	Negative
...	...	...	...	...	...	...	...	...	...
6387	BITID200403	Rahaman	21	M	1715678161	Begomgonj	29.12.2020	31.12.2020	Negative
6388	BITID2004083	Tuhi	22	F	1794744559	hatiya, Noakhali	29.12.2020	31.12.2020	Negative
6389	BITID2004083	Tasmia	24	F	1863746961	Begomgonj	29.12.2020	31.12.2020	Negative
6390	BITID2004081	Naznin Ferdaous	34	F	1753337756	Durgapur, Noakhali	29.12.2020	31.12.2020	Negative
6391	BITID2004082	Nila Akter	25	F	1753337756	Kabirhat	29.12.2020	31.12.2020	Positive

6392 rows × 10 columns

b) Shape of the dataset

Simply, the shape tuple will give us the dimensions of the dataset.

```
In [27]: df.shape
Out[27]: (6392, 9)
```

c) Data Analysis for Gender

Data analysis of the Covid 19 affected rate of male and female. That means women will be more affected or men will be more affected.

d) Data Analysis for Gender

Data analysis for the patients result based on whole Noakhali District and confirmed cases. It means the result will be more positive or the result will be more negative.

3.4 DEVELOPED MODEL

This is most important phase which includes model building for prediction of performance. In this we have implemented three machine learning algorithms for prediction. These algorithms include Random Forest, Support Vector Classifier and K-Nearest Neighbor.

3.4.1 Random Forest

Random forest is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most used algorithms, because of its simplicity and diversity (it can be used for both classification and regression tasks). In this post we'll learn how the random forest algorithm works, how it differs from other algorithms and how to use it

One big advantage of random forest is that it can be used for both classification and regression problems, which form the majority of current machine learning systems. Let's look at random forest in classification, since classification is sometimes considered the building block of machine learning. Below you can see how a random forest would look like with two trees:

Random forest has nearly the same hyper parameters as a decision tree or a bagging classifier. Fortunately, there's no need to combine a decision tree with a bagging classifier because you can easily use the classifier-class of random forest. With random forest, you can also deal with regression tasks by using the algorithm's regressor.
Random forest adds additional randomness to the model, while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model.
Therefore, in random forest, only a random subset of the features is taken into consideration by the algorithm for splitting a node. You can even make

trees more random by additionally using random thresholds for each feature rather than searching for the best possible thresholds (like a normal decision tree does).

3.4.2 K-Nearest Neighbor Model

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm. K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems. K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data. K-Nearest Neighbour (KNN) is one of the simplest Machine Learning algorithms based on Supervised Learning technique.

K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K-NN algorithm. K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems. K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.KNN algorithm at

the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

(a) Working Principle

Suppose there are two categories, i.e., Category A and Category B, and we have a new data point x1, so this data point will lie in which of these categories. To solve this type of problem, we need a K-NN algorithm. With the help of K-NN, we can easily identify the category or class of a particular dataset. Consider the below diagram:

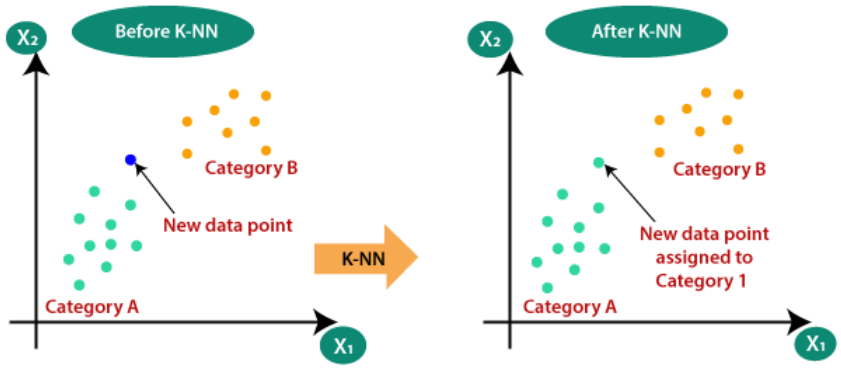

Figure 3.13: KNN with a class or category

Suppose we have a new data point and we need to put it in the required category. Consider the below image:

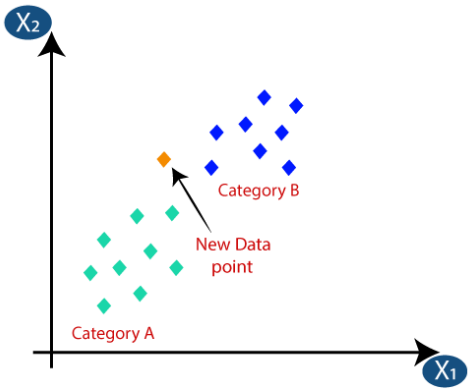

Figure 3.13: KNN with a new data point

Firstly, we will choose the number of neighbors, so we will choose the k=5.Next, we will calculate the Euclidean distance between the data points. The Euclidean distance is the distance between two points, which we have already studied in geometry.

By calculating the Euclidean distance we got the nearest neighbors, as three nearest neighbors in category A and two nearest neighbors in category B. Consider the below image:

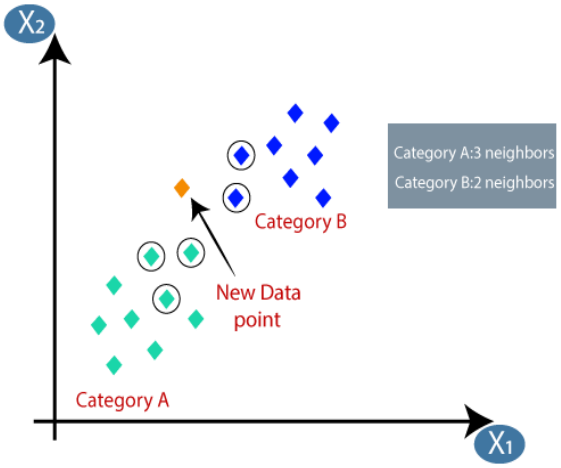

Figure 3.16: A new data point close to a class

As we can see the 3 nearest neighbors are from category A, hence this new data point must belong to category A.

(b) SelectingK value

- There is no particular way to determine the best value for "K", so we need to try some values to find the best out of them. The most preferred value for K is 5.
- A very low value for K such as K=1 or K=2, can be noisy and lead to the effects of outliers in the model.
- Large values for K are good, but it may find some difficulties.

(a) Algorithm

The K-NN working can be explained on the basis of the below algorithm:

- Step-1: Select the number K of the neighbors
- Step-2: Calculate the Euclidean distance of K number of neighbors
- Step-3: Take the K nearest neighbors as per the calculated Euclidean distance.
- Step-4: Among these k neighbors, count the number of the data points in each category.
- Step-5: Assign the new data points to that category for which the number of the neighbor is maximum.

- Step-6: Our model is ready.

3.4.3 Support Vector Machine Model

"Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well .

In the SVM algorithm, we are looking to maximize the margin between the data points and the hyperplane. The loss function that helps maximize the margin is hinge loss.

SVM is a famous supervised machine learning algorithm used for classification as well as regression algorithms. However, mostly it is preferred for classification algorithms. It basically separates different target classes in a hyperplane in n-dimensional or multidimensional space. The main motive of the SVM is to create the best decision boundary that can separate two or more classes(with maximum margin) so that we can correctly put new data points in the correct class.

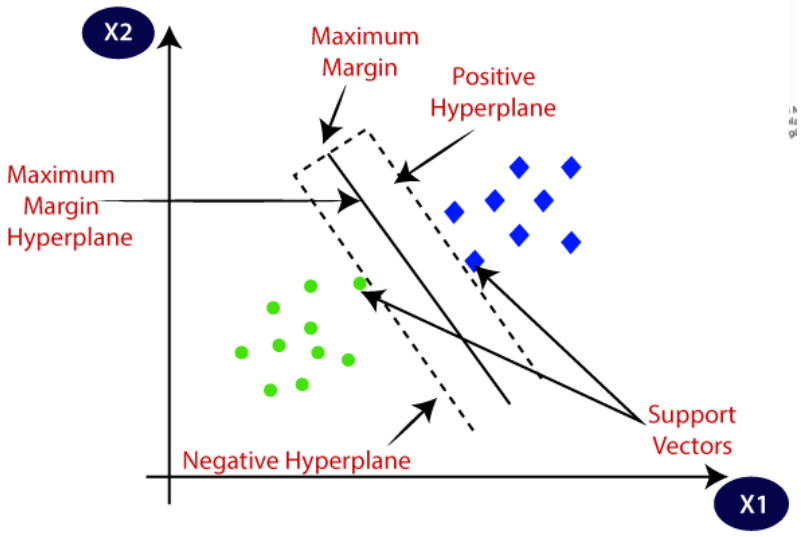

Fig 3.4.3 : Support vector machine

Algorithm

Although the SVM can be applied to various optimization problems such as regression, the classic problem is that of data classification. The basic idea is shown in figure . The data points are identified as being positive or negative, and the problem is to find a hyperplane that separates the data points by a maximal margin. The above figure shows the 2-dimensional case where the data points are linearly separable. The mathematics of the problem to be solved is the following:

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2}\|w\|, \\ \text{s.t} \quad & y_i = +1 \rightarrow \vec{w} \cdot \vec{x_i} + b \geq +1 \\ & y_i = -1 \rightarrow \vec{w} \cdot \vec{x_i} - b \leq -1 \\ & \text{s.t} \quad y_i(\vec{w} \cdot \vec{x_i} + b) \geq 1, \forall_i \end{aligned} \qquad \text{(i)}$$

The identification of the each data point xi is yi, which can take a value of +1 or (representing positive or negative respectively). The solution hyper-plane is the following:

$$u = \vec{w} \cdot \vec{x} + b \quad (2)$$

The scalar b is also termed the bias. A standard method to solve this problem is to apply the theory of Lagrange to convert it to a dual Lagrangian problem. The dual problem is the following:

$$\begin{aligned} \min_{\alpha} \quad & ¥(\vec{\alpha}) = \min_{\alpha} \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} y_i y_j \left(\vec{x_i} \cdot \vec{x_j}\right) \alpha_i \alpha_j - \sum_{i=1}^{N} \alpha_i \\ & \sum_{i=1}^{N} \alpha_i y_i = 0 \\ & \alpha_i \geq 0, \quad \forall_i \end{aligned} \qquad \text{(ii)}$$

3.5 PERFORMANCE ANALYSIS

Performance analysis involves systematic observations to enhance performance that improves decision making of data and gives feedback to our system. It is the process of studying or evaluating the performance of a particular scenario in comparison of the objective which to be achieved. For analyses we use confusion matrix, accuracy, precision, recall and f1 score

(a) Confusion Matrix

A Confusion matrix is an N × N matrix used for evaluating the performance of a classification model, where N is the number of target classes. The matrix compares the actual target values with those predicted by the machine learning model. This gives us a holistic view of how well our classification model is performing and what kinds of errors it is making. For a binary classification problem, we would have a 2×2 matrix as shown below with 4 values:

	Actual Value (positive)	Actual Value (Negative)
Predicted Value (positive)	True Positive (TP)	True Negative (TN)
Predicted Value (negative)	False positive (FP)	False Negative (FN)

Figure 3.5: Confusion Matrix

The target variable has two values: Positive or Negative The columns represent the actual values of the target variable The rows represent the predicted values of the target variable

*True Positive (TP) :*The predicted value matches the actual value. The actual value was positive and the model predicted a positive value.

*True Negative (TN) :*The predicted value matches the actual value. That means the actual value was negative and the model predicted a negative value

False Positive (FP) – Type 1 error: The predicted value was falsely predicted. That means the actual value was negative but the model predicted a positive value. Also known as the Type 1 error.

False Negative (FN) – Type 2 error: The predicted value was falsely predicted**.** That means the actual value was positive but the model predicted a negative value**.** Also known as the Type 2 error.

(b) Correlation Matrix

The Correlation matrix is an important data analysis metric that is computed to summarize data to understand the relationship between various variables and make decisions accordingly. It is also an important pre-processing step in Machine Learning pipelines to compute and analyze the correlation matrix where dimensionality reduction is desired on a high-dimension data.

We mentioned how each cell in the correlation matrix is a 'correlation coefficient' between the two variables corresponding to the row and column of the cell. Each row and column represents a variable, and each value in this matrix is the correlation coefficient between the variables represented by the corresponding row and column.

(c) Accuracy

In confusion matrix the accuracy is the ratio of correct prediction to the total predictions made. It is often presented as a percentage by multiplying the result by 100. The numerical value of accuracy represents the proportion of true positive results (both true positive and true negative) in the selected population. The formula shown as below

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{TN}+\text{FN}} \qquad \text{(i)}$$

(d) Precision

Precision tells us how many of the correctly predicted cases actually turned out to be positive. This would determine whether our model is reliable or not. Precision is a useful metric in cases where False Positive is a higher concern than False Negatives.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \qquad \textbf{(ii)}$$

(e) Recall

Recall tells us how many of the actual positive cases we were able to predict correctly with our model. Recall is a useful metric in cases where False Negative trumps False Positive.

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \qquad \text{(iii)}$$

(f) F1-Score

In practice, when we try to increase the precision of our model, the recall goes down, and vice-versa. The F1-score captures both the trends in a single value:F1-score is a harmonic mean of Precision and Recall, and so it gives a combined idea about these two metrics. It is maximum when Precision is equal to Recall. But there is a catch here. The interpretability of the F1-score is poor.

$$\text{F1 Score} = \frac{2}{\frac{1}{Recall} + \frac{1}{Precision}} \qquad \text{(iv)}$$

CHAPTER 4

RESULT AND DISCUSSION

4.1 INTRODUCTION

In this section, we used different regression and classification algorithms for calculation of performance evaluation metrics. Results are produced from all of the used algorithms. This section presents the numerical analysis performed with random forest, k nearest neighbor and support vector machine. Compare the results obtained by various performance metrics and finally conclude the obtained results by comparison. We evaluate data by graph, correlation matrix, confusion matrix. . After performing various calculation the final result is obtained by the algorithms output. The following section described the results briefly.

4.2 DATA EVALUATION

The data evaluation is done by using correlation plot. It is a 2D plot that focused on the statistics. The correlation plot will provide the fundamental concept of the feature variables and how the features are reacting in the dataset and pointing out the key features. The data visualization chooses essential features for training and produces the correct output. A correlation matrix is a table showing correlation coefficients between variables, each cell in the table shows the correlation between two variables. Correlation plot denotes how changes between two variables relate. Two variables that change in the same direction are positively correlated. A change in opposite directions implies negative correlation. The positive signed values are positively correlated and negative signed values are negatively correlated to the target variable. From this figure we can determine the variables which have a good impact on final outcome and thus analyses will also improved.

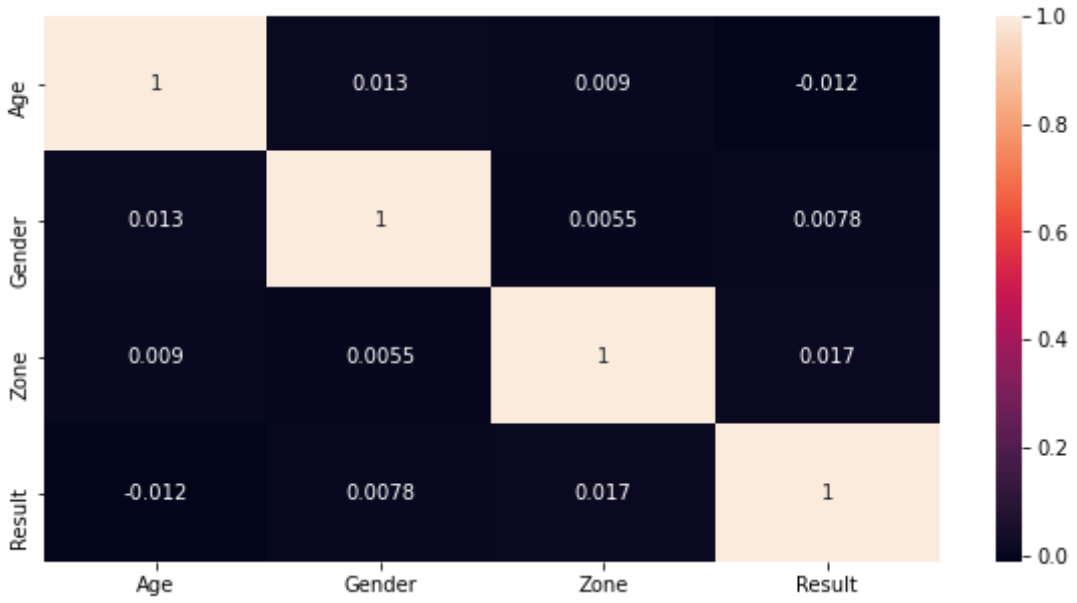

Fig 4.1: The correlation matrix plot for the important feature

4.2.1 Data Evaluation for Gender

This graph gives us an idea of how many observations each bin holds, this is a good way to visualize data. The shapes of the bins tell us whether an attribute is Gaussian, skewed, or has an exponential distribution. It also hints us about outliers.

Data analysis of the Covid 19 affected rate of male and female. That means women will be more affected or men will be more affected.

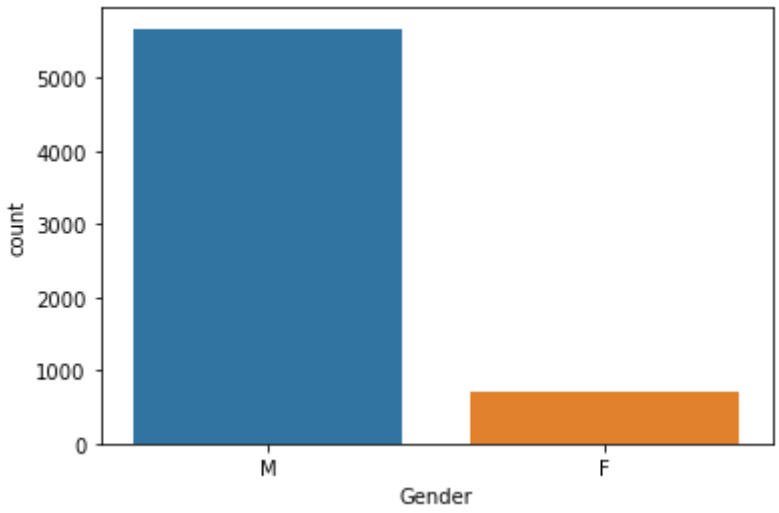

Fig 4.2.1: Data Evaluation for Gender Based on Dataset

4.2.2 Data Evaluation for Result

This graph gives us an idea of how many observations each bin holds, this is a good way to visualize data .The shapes of the bins tell us whether an attribute is Gaussian, skewed, or has an exponential distribution. It also hints us about outliers. Data analysis for the patients result based on whole Noakhali District and confirmed cases. It means the result will be more positive or the result will be more negative.

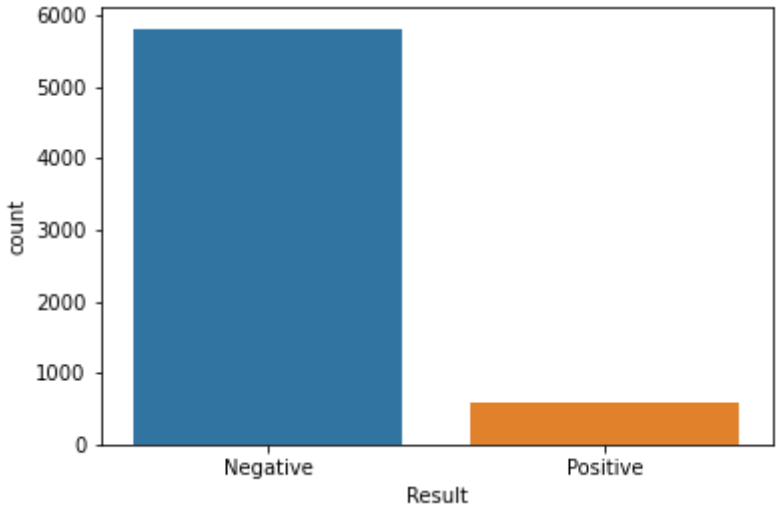

Fig 4.2.2: Data Evaluation for Result Based on Dataset

4.3 PERFORMANCE ANALYSIS

There are so many performance analysis measures when it comes to selecting a classification model. Such as, confusion matrix, classification report. And the classification report has some features like precision, recall,F1- score, Support, accuracy, micro average, weighted average.

4.3.1 Random Forest

We are performing random forest classification for two classes (0 or 1), and we often get confusion matrix to evaluate the model. Element [0,0] is the number of correct classifications in the 0 class and element [1,1] is the number of correct classifications in the 1 class.

In this type of algorithm, different type of models or identical models are merged together to implement Random Forest Algorithm (SE Hamby, 2008).Merging of different algorithm

creates random forest. This algorithm is used for regression and classification. But this algorithm has more application in class.

(a) *Confusion Matrix*

	Actual Value (positive)	Actual Value (Negative)
Predicted Value (positive)	841	18
Predicted Value (negative)	97	3

Fig 4.3.1 : Confusion Matrix For Random Forest

In the above figure for the true positive(TP) value is 841 which means the predicted value matches the actual value. The actual value was positive and the model predicted a positive value. The true negative(TN) value is 18 which means the predicted value matches the actual value. The actual value was negative and the model predicted a negative value. The false positive(FP) value is 97 this is also known as type 1 error which means the predicted value was falsely predicted. The actual value was negative but the model predicted a positive value. The false negative(FN) value is 3 which means The predicted value was falsely predicted. The actual value was positive but the model predicted a negative value.

b) Classification report

	Precision	Recall	F1 Score	Support
0	0.90	0.98	0.94	859
1	0.14	0.03	0.05	100
Accuracy			0.88	959
Macro Average	0.52	0.50	0.49	959
Weighted Average	0.82	0.88	0.84	959

Fig 4.3.2 : Classification Report for Random Forest

In Random Forest the accuracy is calculated from the equation (TP+TN)/(TP+FP+TN+FN) which is 96%.The TP(true positive), TN(true negative), FP(false positive), FN(false negative) values are obtained from the confusion matrix. The equation TP/(TP+FP) is used for precision calculation. The precision value is 100%. It tells us how many of the correctly predicted cases actually turned out to be positive. This would determine whether our model is reliable or not. Precision is a useful metric in cases where False Positive is a higher concern than False Negatives. The equation TP/(TP+FN) for recall value is 94% which tells us how many of the actual positive cases we were able to predict correctly with our model. Recall is a useful metric in cases where false negative trumps false positive. And the equation for finding the value of F1 score is calculated by 2/((1/recall)+(1/precision)) is 89%. In practice, when we try to increase the precision of our model, the recall goes down, and vice-versa. The F1-score captures both the trends in a single value. F1-score is a harmonic mean of Precision and Recall, and so it gives a combined idea about these two metrics. It is maximum when Precision is equal to Recall. But there is a catch here. The interpretability of the F1-score is poor. Now in the same manner the other two algorithms (KNN and SVM) predicted their evaluation metrices which is shown in the table.

d) ROC Curve

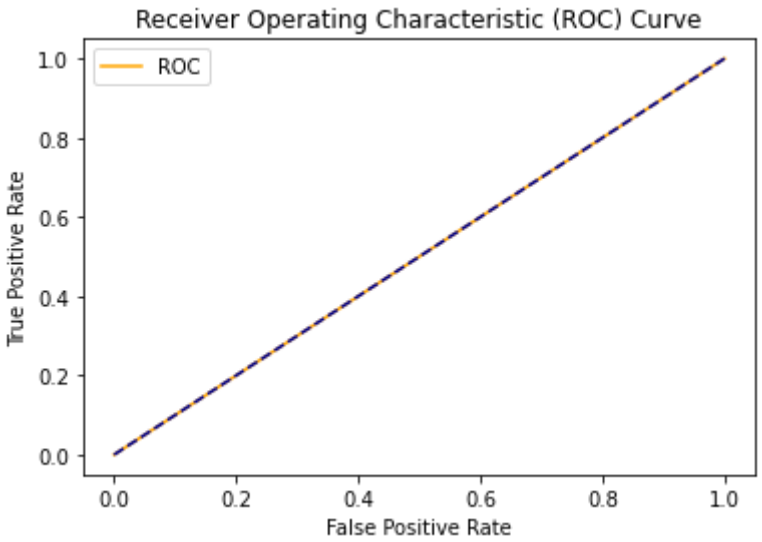

Fig 4.3.3 : ROC Curve For Random Forest

ROC curve is a graph showing the performance of a classification model and all classification thresholds. This curve plots two parameters : True Positive Rate and False Positive Rate. The true positive rate is also known as sensitivity. False positive rate is also known as probability of false alarm. In the above figure indicates that fall under the area at the top left corner indicate good performance, and ROC curve fall in the other area at the bottom right corner indicate poor performance levels. We have only two points here (0,0) and (1,1) . Claiming that we have a straight line that is not very satisfying.

d) Accuracy

The accuracy for Random Forest is that we got from the dataset is 88%. That means the Random Forest algorithm-generated good accuracy rate.

4.4.4 K Nearest Neighbor

We are performing K Nearest Neighbor classification for two classes (0 or 1), and we often get confusion matrix to evaluate the model. Element [0,0] is the number of correct classifications in the 0 class and element [1,1] is the number of correct classifications in the 1 class.

It is one of simplest type of machine learning algorithm and it is based on supervised learning. This algorithm predicts the occurrence of the dependent variable. Binary value is used to indicate the dependent variable. This algorithm predicts and calculate the success rate of probability.

(a) *Confusion Matrix*

	Actual Value (positive)	Actual Value (Negative)
Predicted Value (positive)	830	29
Predicted Value (negative)	99	1

Fig 4.4.1 : Confusion Matrix For Random Forest

In the above figure for the true positive (TP) value is 830 which means the predicted value matches the actual value. The actual value was positive and the model predicted a positive value. The true negative(TN) value is 29 which means the predicted value matches the actual value. The actual value was negative and the model predicted a negative value. The false positive(FP) value is 99 this is also known as type 1 error which means the predicted value was falsely predicted. The actual value was negative but the model predicted a positive value. The false negative(FN) value is 1 which means The predicted value was falsely predicted. The actual value was positive but the model predicted a negative value.

b) Classification report

	Precision	Recall	F1 Score	Support
0	0.89	0.97	0.93	859
1	0.03	0.01	0.02	100
Accuracy			0.87	959
Macro Average	0.46	0.49	0.47	959
Weighted Average	0.80	0.87	0.83	959

Fig 4.4.2 : Classification Report for K Nearest Neighbor

In K Nearest Neighbor the accuracy is calculated from the equation (TP+TN)/(TP+FP+TN+FN) which is 95%.The TP(true positive), TN(true negative), FP(false positive), FN(false negative) values are obtained from the confusion matrix. The equation TP/(TP+FP) is used for precision calculation. It tells us how many of the correctly predicted cases actually turned out to be positive. This would determine whether our model is reliable or not. Precision is a useful metric in cases where False Positive is a higher concern than False Negatives. The equation TP/(TP+FN) for recall value is 94% which tells us how many of the actual positive cases we were able to predict correctly with our model. Recall is a useful metric in cases where false negative trumps false positive. And the equation for finding the value of F1 score is calculated by 2/((1/recall)+(1/precision)) is 87%. In practice, when we try to increase the precision of our model, the recall goes down, and vice-versa. The F1-score captures both the trends in a single value. F1-score is a harmonic mean of Precision and Recall,

and so it gives a combined idea about these two metrics. It is maximum when Precision is equal to Recall. But there is a catch here.

c) ROC Curve

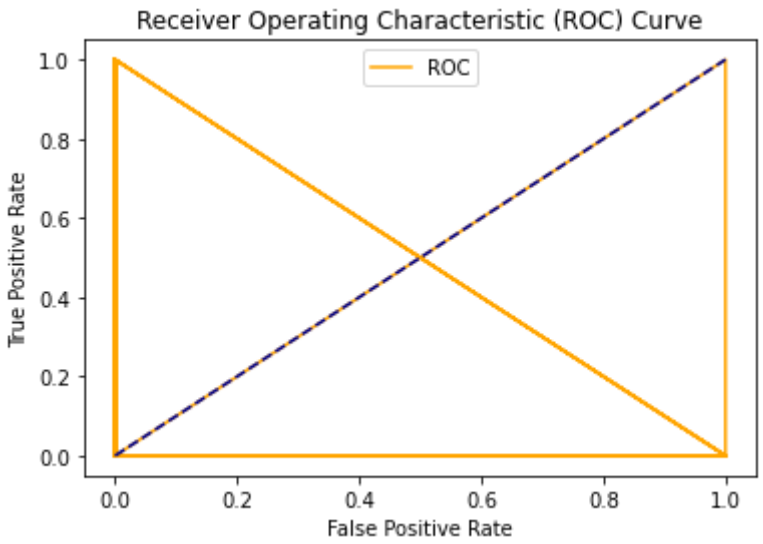

Fig 4.4.3 : ROC Curve For K Nearest Neighbor

This figure indicates that there's no better odds of defecting something. The closer the ROC curve get to the top left corner the better the test is overall. So in this curve we have a perfect classifier at the top left corner.

d) Accuracy

The accuracy for K Nearest Neighbor is that we got from the dataset is 86.65%. That means the K Nearest Neighbor algorithm-generated lower accuracy rate than Random Forest.

4.4.5 Support Vector Machine

We are performing K Nearest Neighbor classification for two classes (0 or 1), and we often get confusion matrix to evaluate the model.

In machine learning, Support vector machines are supervised learning models with associated learning algorithms that analyze data for classification and regression analysis.

(a) *Confusion Matrix*

	Actual Value (positive)	Actual Value (Negative)
Predicted Value (positive)	859	0
Predicted Value (negative)	99	0

Fig 4.5.1 : Confusion Matrix For SVM

In the above figure for the true positive (TP) value is 859 which means the predicted value matches the actual value. The actual value was positive and the model predicted a positive value. The true negative(TN) value is 0 which means the predicted value matches the actual value. The actual value was negative and the model predicted a negative value. The false positive(FP) value is 99 this is also known as type 0 error which means the predicted value was falsely predicted. The actual value was negative but the model predicted a positive value. The false negative(FN) value is 1 which means The predicted value was falsely predicted. The actual value was positive but the model predicted a negative value.

b) Classification report

	Precision	Recall	F1 Score	Support
0	0.90	1.00	0.94	859
1	0.00	0.00	0.01	100
Accuracy			0.90	959
Macro Average	0.45	0.50	0.47	959
Weighted Average	0.80	0.90	0.85	959

Fig 4.5.2 : Classification Report For SVM

In Support vector machine the accuracy is calculated from the equation (TP+TN)/(TP+FP+TN+FN) which is 95%.The TP(true positive), TN(true negative), FP(false positive), FN(false negative) values are obtained from the confusion matrix. The equation TP/(TP+FP) is used for precision calculation. It tells us how many of the correctly predicted cases actually turned out to be positive. This would determine whether our model is reliable or not. Precision is a useful metric in cases where False Positive is a higher concern than False Negatives. The equation TP/(TP+FN) for recall value is 94% which tells us how many of the actual positive cases we were able to predict correctly with our model. Recall is a useful metric in cases where false negative trumps false positive. And the equation for finding the value of F1 score is calculated by 2/((1/recall)+(1/precision)) is 87%. In practice, when we try to increase the precision of our model, the recall goes down, and vice-versa. The F1-score captures both the trends in a single value. F1-score is a harmonic mean of Precision and Recall, and so it gives a combined idea about these two metrics. It is maximum when Precision is equal to Recall. But there is a catch here.

c) ROC Curve

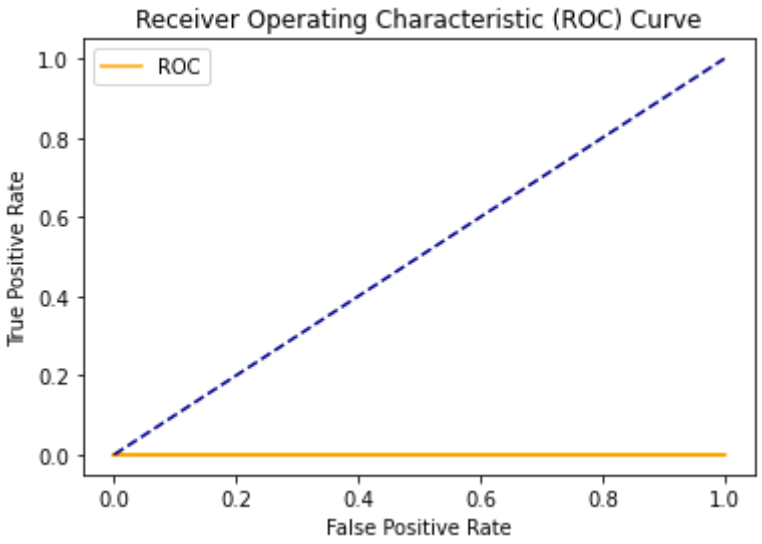

Fig 4.5.3 : ROC Curve For SVM

In the above figure indicates that fall under the area at the top left corner indicate good performance, and ROC curve fall in the other area at the bottom right corner indicate good performance levels also. So we can have a straight line that is very satisfying.

d) Accuracy

The accuracy for Support Vector Machine is that we got from the dataset is 89.58% That means the Support Vector Machine algorithm-generated highest accuracy rate than among Random Forest and K Nearest Neighbor.

4.4.6 Comparison of Accuracy Rate among the Algorithms:

Techniques	Random Forest	K Nearest Neighbor	Support Vector Machine
Accuracy Rate	88%	86.65%	89.58%.

Fig 4.5 6: Comparison of Accuracy Rate among the Algorithms

From the observation, it is found that the Support Vector Machine -generated maximum accuracy rate. Machine learning is used to classify the model based on certain features and aim to generate 100% accuracy.

CHAPTER 5

FUTURE WORKS AND CONCLUSION

5.1 Future Works

First, understanding the propensity of these viruses to jump between species, to establish infection in a new host, and to identify significant reservoirs of corona viruses will dramatically aid in our ability to predict when and where potential epidemics may occur. These studies should lead to a large increase in the number of suitable therapeutic targets to combat infections. As bats seem to be a significant reservoir for these viruses, it will be interesting to determine how they seem to avoid clinically evident disease and become persistently infected. Second, many of the non-structural and accessory proteins encoded by these viruses remain uncharacterized with no known function, and it will be important to identify mechanisms of action for these proteins as well as defining their role in viral replication and pathogenesis.

5.2 Conclusion

Corona virus is a new threat for human civilization. Generally virus born diseases in most of cases do not appear to be lethal except some specific cases including excessive vulnerability and anti drug resistance from unregulated application of antibiotics.It is likely that these viruses will continue to emerge and to evolve and cause both human and veterinary outbreaks owing to their ability to recombine, mutate, and infect multiple species and cell types. There are hundreds of corona viruses, most of which circulate in animals. Only seven of these viruses infect humans and four of them cause symptoms of the common cold. Already this virus has affected over the world. SARS, a beta corona virus emerged in 2002 and was controlled mainly by aggressive public health measures. There have been no new cases since 2004. MERS emerged in 2012, still exists in camels, and can infect people who have close contact with them

But, three times in the last 20 years, a corona virus has jumped from animals to humans to cause severe disease.

.

REFERENCES

[1] W[oo PC, Huang Y, Lau SK, Yuen KY. Coronavirus genomics and bioinformatics analysis. Viruses, 2010; 2: 1804-20.

[2] Drexler, J.F., Gloza-Rausch, F., Glende, J., Corman, V.M., Muth,D., Goettsche, M., Seebens, A., Niedrig, M., Pfefferle, S., Yor-danov, S., Zhelyazkov, L., Hermanns, U., Vallo,

[3] Lukashev, A.,Muller, M.A., Deng, H., Herrler, G., Drosten, C., Genomic characterization of severe acute respiratory syndrome-related coronavirus in European bats and classification of coronaviruses based on partial RNA-dependent RNA polymerase gene sequences. J. Virol, 2010; 84: 11336–11349.

[4] Yin, Y., Wunderink, R. G. MERS, SARS and other coronaviruses as causes of pneumonia. Respirology, 2018; 23(2): 130-137.

[5] Peiris, J. S. M., Lai S. T., Poon L. et. al. Coronavirus as a possible cause of severe acute respiratory syndrome. The Lancet, 2003; 361(9366): 1319-1325.

[6] Zaki AM, van Boheemen S, Bestebroer TM, Osterhaus AD,Fouchier RA. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. N. Engl. J. Med, 2012; 367: 1814–20. Seven days in medicine: 8-14 Jan 2020. BMJ, 2020; 368-132.31948945.

[7] Imperial College London. Report 2: estimating the potential total number of novel coronavirus cases in Wuhan City, China. Jan. disease-analysis/news--wuhan-coronavirus, 2020.

[8] European Centre for Disease Prevention and Control data.Geographical distribution of 2019- nCov cases. Available online: (https://www.ecdc.europa.eu/en/geographical-distribution-2019-ncov-cases) (accessed on 05 February 2020).

[9] World Helath Organization, nCoV Situation Report-22 on 12 February, 2020. source/coronaviruse /situation-reports/, 2019.

[10] Gralinski L.; Menachery V; Return of the Coronavirus: 2019- nCoV, Viruses, 2020; 12(2): 135.

[11] Chen Z.; Zhang W.; Lu Y et. al.. From SARS-CoV to Wuhan 2019-nCoV Outbreak: Similarity of Early Epidemic and Prediction of Future Trends.: Cell Press, 2020.

[12] Luk H. K., Li X., Fung J., Lau S. K., Woo P. C. (Molecular epidemiology, evolution and phylogeny of SARS coronavirus. Infection, Genetics and Evolution, 2019; 71: 21-30.

[13] Coronavirinae in ViralZone. expasy.org/785 (accessed on 05 February 2019).

[14] Subissi, L.; Posthuma, C.C.; Collet, A.; Zevenhoven-Dobbe, J.C.; Gorbalenya, A.E.; Decroly, E.; Snijder, E.J.; Canard, B.; Imbert, I.One severe acute respiratory syndrome coronavirus proteincomplex integrates processive RNA polymerase and exonuclease activities. Proc. Natl. Acad. Sci. USA 2014, 111, E3900–E3909.

[15] Zhao L, Jha BK, Wu A, Elliott R, Ziebuhr J, Gorbalenya AE, Silverman RH, Weiss SR. Antagonism of the interferon-induced OAS-RNase L pathway by murine coronavirus ns2 protein is required for virus replication and liver pathology. Cell host & microbe, 2012; 11(6): 607–616.

[16] Barcena M, Oostergetel GT, Bartelink W, Faas FG, Verkleij A, Rottier PJ, Koster AJ, Bos BJ. Cryo-electron tomography of mouse hepatitis virus: Insights into the structure of the coronavirion. Proceedings of the National Academy of Sciences of the United States of America, 2009; 106(2): 582–587.

[13] Neuman BW, Adair BD, Yoshioka C, Quispe JD, Orca G, Kuhn P, Milligan RA, Yeager M, Bucheier MJ. Supramolecular architecture of severe acute
respiratory syndrome coronavirus revealed by electron cryomicroscopy. Journal of virology, 2006; 80(16): 7918–7928.

[14] Beniac DR, Andonov A, Grudeski E, Booth TF. Architecture of the SARS coronavirus prefus on spike. Nature structural & molecular biology, 2006; 13(8): 751–752.

[15] Delmas B, Laude H. Assembly of coronavirus spike protein into trimers and its role in epitope expression. Journal of virology, 1990; 64(11): 5367–5375.

[17] Bosch BJ, van der Zee R, de Haan CA, Rottier PJ. The coronavirus spike protein is a class I virus fusion protein: structural and functional characterization of the fusion core complex. J Virol, 2003; 77(16): 8801–8811.

[18] Collins AR, Knobler RL, Powell H, Buchmeier MJ. Monoclonal antibodies to murine hepattis virus-4 (strain JHM) define the viral glycoprotein responsible for attachment and cell--cell fusion. Virology, 1982; 119(2): 358–371.

[19] Abraham S, Kienzle TE, Lapps W, Brian DA. Deduced sequence of the bovine coronavirus spike protein and identification of the internal proteolytic cleavage site. Virology, 1990; 176(1): 296–301.

[20] Luytjes W, Sturman LS, Bredenbeek PJ, Charite J, van der Zeijst BA, Horzinek MC, Spaan WJ. Primary structure of the glycoprotein E2 of coronavirus MHV-A59 and identification of the trypsin cleavage site. Virology, 1987; 161(2): 479–487.
[21]. De Groot RJ, Luytjes W, Horzinek MC, van der Zeijst BA, Spaan WJ, Lenstra JA. Evidence for a coiled-coil structure in the spike proteins of coronaviruses. J Mol Biol, 1987; 196(4): 963–966.
[22] Armstrong J, Niemann H, Smeekens S, Rottier P, Warren G. Sequence and topology of a model intracellular membrane protein, E1 glycoprotein, from a coronavirus. Nature, 1984; 308(5961): 751–752.
[23] Nal B, Chan C, Kien F, Siu L, Tse J, Chu K, Kam J, Staropoli I, Crescenzo-Chaigne B, Escriou N, van der Werf S, Yuen KY, Altmeyer R. Differential maturation and subcellular localization of severe acute respiratory syndrome coronavirus surface proteins S, M and E. The Journal of general virology, 2005; 86(Pt 5): 1423–1434.
[24] Subissi, L.; Posthuma, C.C.; Collet, A.; Zevenhoven-Dobbe, J.C.; Gorbalenya, A.E.; Decroly, E.; Snijder, E.J.; Canard, B.; Imbert, I.One severe acute respiratory syndrome coronavirus proteincomplex integrates processive RNA polymerase and exonuclease activities. Proc. Natl. Acad. Sci. USA 2014, 111, E3900–E3909.
[25] Zhao L, Jha BK, Wu A, Elliott R, Ziebuhr J, Gorbalenya AE, Silverman RH, Weiss SR. Antagonism of the interferon-induced OAS-RNase L pathway by murine coronavirus ns2 protein is required for virus replication and liver pathology. Cell host & microbe, 2012; 11(6): 607–616.
[26] Killerby ME, Biggs HM, Haynes A, Dahl RM,et al. Human coronavirus circulation in the United States 2014 – 2017external icon.Journal of Clinical Virology. Vol 101; 2018 Apr; 101:52-6
[27] Ashikujaman Syed. (2019). Ebola VirusDisease. Int. J. Curr. Res. Med. Sci. 5(3): 18-
[28] Content source: National Center forImmunization and Respiratory Diseases(NCIRD).

[29] Coronaviruses, including SARS-CoV (RedBook, American Academy of Pediatrics,2018)external icon
[30] Severe Acute Respiratory Syndrome (SARS),CDC
[31] Severe Acute Respiratory Syndrome (SARS),WHOexternal icon
[32] Yin, Y., Wunderink, R. G. MERS, SARS and other coronaviruses as causes of pneumonia. Respirology, 2018; 23(2): 130-137.

APPENDIX

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
df=pd.read_csv("New dataset.csv")
df
df.shape
df.dtypes
sns.countplot(df["Result"],label="count")
sns.countplot(df["Sex"],label="count")
sns.countplot(df["Age"],label="count")
sns.countplot(df["Address"],label="count")
df=df.drop(['Lab ID','Name','Contact Number','Date of Sample collection','Date of
Test'],axis=1)
df
```

```
from sklearn.preprocessing import LabelEncoder

lb = LabelEncoder()
df['Age']=lb.fit_transform(df['Age'])
df['Sex']=lb.fit_transform(df['Sex'])
df['Address']=lb.fit_transform(df['Address'])
df['Result']=lb.fit_transform(df['Result'])
df
from sklearn.preprocessing import LabelEncoder

lb = LabelEncoder()
df['Age']=lb.fit_transform(df['Age'])
df['Sex']=lb.fit_transform(df['Sex'])
df['Address']=lb.fit_transform(df['Address'])
df['Result']=lb.fit_transform(df['Result'])
df
plt.figure(figsize=(10,5))

correlation_matrix=df.corr()

sns.heatmap(correlation_matrix,annot=True)
x=df.iloc[:,0:3].values
x
y=df.iloc[:,-2].values
y
from sklearn.model_selection import train_test_split
```

```
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=.15,random_state=0)
print ('Train set:', x_train.shape,  y_train.shape)
print ('Test set:', x_test.shape,  y_test.shape)
from sklearn.metrics import accuracy_score
from sklearn.metrics import confusion_matrix
from sklearn.metrics import classification_report
from sklearn.metrics import f1_score
import warnings
warnings.filterwarnings("ignore")
from sklearn.preprocessing import StandardScaler

sc = StandardScaler()
x_train = sc.fit_transform(x_train)
x_test = sc.transform(x_test)
print("random forest")
from sklearn.ensemble import RandomForestClassifier

rfc=RandomForestClassifier(n_estimators=100)

rfc.fit(x_train,y_train)
y_pred=rfc.predict(x_test)

print("f1 score:",f1_score(y_test, y_pred,average="micro"))
print("accuracy score:",accuracy_score(y_test, y_pred))

print("confusion matrix:")
print(confusion_matrix(y_test, y_pred))
```

```
print("classification report")
print(classification_report(y_test,y_pred,labels=np.unique(y_pred)))
#y_pred=y_pred[:,1]
ns_fpr, ns_tpr, _ = roc_curve(y_test, y_pred)
plt.plot(ns_fpr, ns_tpr, color='orange', label='ROC')
plt.plot([0, 1], [0, 1], color='darkblue', linestyle='--')
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver Operating Characteristic (ROC) Curve')
plt.legend()
plt.show()
print("KNearestNeighbor ")
from sklearn.linear_model import KNearestNeighbor

regression = KNearestNeighbor(solver = 'liblinear', random_state = 0)
regression.fit(x_train, y_train)

y_pred = regression.predict(x_test)
print("accuracy score:", accuracy_score(y_test, y_pred))
print("confustion matrix:\n",confusion_matrix(y_test,y_pred))
print("f1 score:",f1_score(y_test,y_pred,average="macro"))
print("classification report:")
print(classification_report(y_test,y_pred))
from sklearn.naive_bayes import GaussianNB
gnb = GaussianNB()
gnb.fit(x_train, y_train)
y_pred = gnb.predict(x_test)
```

```
print("accuracy score:", accuracy_score(y_test, y_pred))
print("confustion matrix:\n",confusion_matrix(y_test,y_pred))
print("f1 score:",f1_score(y_test,y_pred,average="micro"))
print("classification report:",classification_report(y_test,y_pred))
ns_fpr, ns_tpr, _ = roc_curve(y_test, y_pred)

plt.plot(y_test, y_pred, color='orange', label='ROC')
plt.plot([0, 1], [0, 1], color='darkblue', linestyle='--')
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver Operating Characteristic (ROC) Curve')
plt.legend()
plt.show()
print("Support Vector Machine")
from sklearn.svm import SVC
support = SVC(kernel='rbf', random_state = 1)
support.fit(x_train, y_train)
y_pred = support.predict(x_test)
print("accuracy score:", accuracy_score(y_test, y_pred))
print("confustion matrix:\n",confusion_matrix(y_test,y_pred))
print("f1 score:",f1_score(y_test,y_pred,average="micro"))
print("classification report:")
print(classification_report(y_test,y_pred))
#y_pred=y_pred[:,1]
ns_fpr, ns_tpr, _ = roc_curve(y_test, y_pred)
plt.plot(ns_fpr, ns_tpr, color='orange', label='ROC')
plt.plot([0, 1], [0, 1], color='darkblue', linestyle='--')
```

```
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver Operating Characteristic (ROC) Curve')
plt.legend()
plt.show()
```