Prediction of Mortality Rate among Various Age Groups of Covid-19 Patients in India Using Naive Bayes Algorithm for Effective Decision Making

Sushree Priyadarsini Das, Murali Krishna Senapaty

Dept of Computer Science and Engineering, GIET University, Gunupur
Email: priyadarsini.das044@gmail.com, muralisenapaty@giet.edu

ABSTRACT : The noble corona virus or COVID-19 has spread in almost 213countries across the globe and its wrath has affected more than 7 million people and has claimed about 429067 lives by june 1st week. Our country, India stands at the 4th position among the affected countries worldwide. In India only there are more than 3 lakhs positive cases of COVID-19 out of which 8895 people lost their precious lives to this deadly disease. In a densely populated country like India, where about 464 people live in 1 km2 of area and about 550 hospital beds are available per 1 million populations, we will soon be run out of beds if the confirmed cases increase at the present rate. So, to save the lives of people the government and hospital personnel must take some decisive measures to avoid the crowd and provide proper healthcare facilities to the infected people. Here we have tried to help the health care worker and hospital staff to make decisions to provide emergency medical facilities to people on a priority basis when the system is in overcrowded condition. It will help them to make quick and right decision to hospitalise the more vulnerable patients with high mortality rates, such as older age group and patients with pre-existing medical conditions. In this model, we have used data set of 27891 confirmed cases of covid-19 people and developed a predictive model using various machine learning algorithms, such as J48 classifier, Naive Bayes, ranker search methods, and Random forestto predict the death rate among the various age group of infected people. The outcome gives around 94% accuracy in predicting the mortality rate.

Keywords: COVID-19, SVM, ROC, Naïve Bayes, Ranker Search, J48, Random Forest

1. INTRODUCTION AND LITERATURE REVIEW

Novel Corona Virus or popularly known as COVID-19 is a highly infectious disease that has brought the world to a halt, affecting 210 countries are 2 international conveyance. The pandemic has become the worst affecting pandemic since World War II. As the virus is a new one and no reliable vaccine has been invented to control the disease so far, thus the virus is spreading and claiming life in an exponential rate. Corona viruses are a group of RNA viruses which belongs to the family Orthocoronavirinae which causes different disease in mammals and birds. In the human it mostly affects the respiratory system. Its more fatal varieties are SARS (Severe Acute Respiratory Syndrome) and MERS(Middle East Respiratory Symptoms)[13].

Corona viruses were first discovered in 1930's when it affected chickens with respiratory tract problem and the morbidity rate was 40-90%. The human corona virus was first discovered in 1960's in United Kingdom and United States. They were isolated in two groups. The virus caused common cold-like symptoms in humans. A Scottish virologist June Almeida of St. Thomas hospital, London was able to show the club like spikes of the virus through the help of electron microscope in the year 1967. Due to its club-like spikes or crown like structure the virus got its name. Hence it was named as Corona virus from Latin corona means crown or wreath.[1][14]

The novel corona virus or covid-19 is an ongoing pandemic which was caused by the group of SARSvirus. The virus came to news in December 2019 and theoutbreak was from a town called Wuhan in China, when some people of that town suffered from an unknown disease that had a flu-like symptom. The doctors were unable to diagnose the disease as it was a new one. In the early stage of the infection, the patients showed no symptoms, which led to the rapid transmission of the virus among other people. It is said that the first patient who suffered from covid-19 got the infection from bats, which is under controversy as some facts say that it was made in a lab in Wuhan, China, and accidentally infected the scientists and hence the common people. The World Health Organisation (WHO) declared the outbreak as a National Health Emergency in 30th January 2020 and a pandemic in 11th march 2020 and was renamed to COVID-19 which was previously known as Novel Corona Virus 2019 [2].The ICTV named the new virus as SARS COV 2(Severe Acute Respiratory Syndrome corona virus 2) due to the

genetically similarity between the corona virus which was responsible for the 2003 SARS outbreak [3][10].

1.1 TRANSMISSION AND INCUBATION PERIOD OF COVID-19

The virus mainly spreads through the close contact people or when a person comes into contact of the infected person. When a person infected with the virus sneeze, coughs, or touches a surface, the virus remains in the environment or on the surface touched. When a healthy individual touch that surface or inhale in that infected environment, the virus enters into the body of the healthy individual and remains active in that body. It takes almost 14 days to show any symptom of infection after the virus enters the body of human beings. As there is no specific treatment is available for the disease only social distancing and maintaining personal hygiene have been made mandatory to control the spread of the virus. This has led almost 80 percent of the world population to stay in lockdown or in seized condition, thus largely affecting the world economy. To avoid further spread of the virus most countries have shut down their international border, travel, and production factories.

1. 2. COVID-19 SCENARIO IN INDIA:

India currently has highest number of confirmed cases in Asia and 4th highest number across the globe. India reported its first COVID-19 confirmed patient on 30th January 2020 in the state of Kerala. The disease was detected in a medical student having travel history from Wuhan, china, the original epic centre of the disease. Then on February 2nd and 3rd two more positive cases detected in the same state with similar travel history. Till the end of February there was very negligible number of cases in the country.

In March 1st week the positive cases again started to increase and India reported its 1st COVID-19 death on 12th march. To combat the pandemic PM Narendra Modi imposed the 1st voluntary lockdown across the country on 22nd march. As the cases began to increase the 1st nationwide lock was imposed on 25th march and it still continues. India crossed the 100k bench march on 9th may, 200k on 3rdJune and 300k on 13thJune. Currently India has 380,532 confirmed cases including 12573 deaths. Recently India's recovered numbers exceeded the number of active cases. India's current fatality rate is 2.80% which is relatively lower than the global fatality rate of 6.13%.

Six cities of India, such as Mumbai, Delhi, Ahmadabad, Chennai, Pune and Kolkata contribute almost 50% of total cases of the country, where Maharastra state has highest number of cases in India. As described earlier, India is densely populated country and people literally get very less space for their living. There are many colonies in the city as well as in the country where a big family consisting of almost 10 members live in a small house. So it becomes very difficult to contain this highly infectious disease from spreading. As this virus is a new one and no reliable treatment is available till now, people who are already infected contaminate other people unknowingly, because most of the time the virus shows no symptoms. This asymptomatic nature of the virus has made it more difficult to stop the spreading and contamination. This has led to community transmission of the disease, because most of the time the patients do not know that they are infected and they are incubating the virus in their body. All these problems and the vast population has led the country to the 4th highest position among the list of infected countries.

Ignorance and lack of proper of information about this new various has created a lot of panic and fear among people. People are panic buying, rather stocking essential items in the fear of total lockdown or unavailability of grocery and food items.

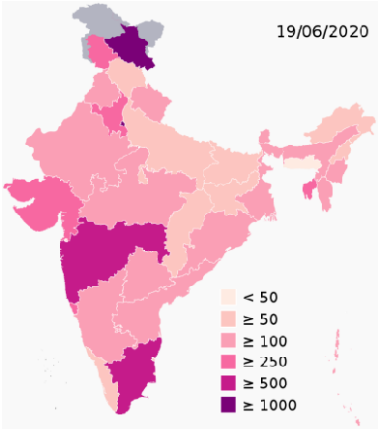


Figure 1: Demographic view of India map showing COVID-19 affected parts of the country

The above figure shows the parts of the country affected by COVID-19[17]. The darker colour represents the most affected part and the lighter ones are for the less affected area of the country. From the picture we can clearly see that, almost all parts of our country is badly affected by the noble corona virus. Epidemiologists have predicted that, India may see the peak of transmission in the coming months and the number of confirmed cases may raise up to 2.5 million, while the number of deaths may top up to 18000.

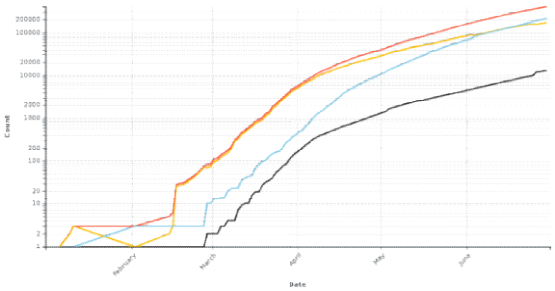


Figure 2: Graph showing the confirmed, recovered, active and deceased cases of COVID-19 in India

Above figure shows the status of COVID-19 patients in India, where the red line is number of total confirmed case, yellow for active cases, blue line is for recoveries

and the black one shows the number of deceased cases [15].

1. 3. HEALTHCARE FACILITY IN INDIA

As discussed above, we know that the present healthcare system in India is in a lamentable state. Here around 550 hospital beds are available per one million populations, which roughly count as 0.5 beds per patient. So from this calculation it is quite clear that not a single medical bed is available for a patient. Below figure shows a tentative graph of hospital facilities in India [16].

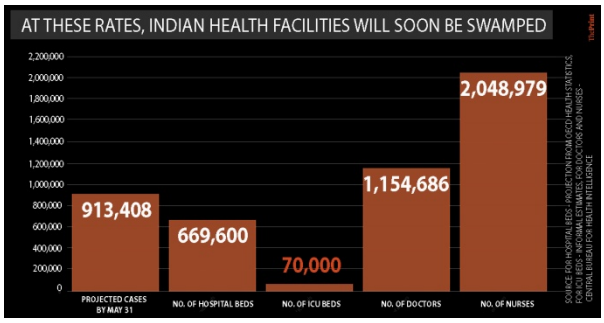


Figure 3: Graph showing available healthcare facility in India

So when a difficult situation arises in future when the hospital is crowded with patients, the hospital management has to take some timely and effective decision to save the precious lives of our citizen. As there are very less number of healthcare workers and medical beds available for treatment of the patients, the management faces a crisis regarding patient admission. To overcome the problem of hospitalisation, here comes the technology, specifically machine learning to the rescue of mankind.

1.4 ROLE OF MACHINE LEARNING IN HEALTHCARE INDUSTRY:

Despite of the fact that, technology is moving fast and making rapid improvements which in turn making the human race less active and the unhealthy lifestyle due to advance technology making people more vulnerable to different lifestyle disease, we cannot deny the fact that, this technology only had helped the healthcare industry to improve rapidly and provide better medical facility to the patients. Fatima Paruk, CMO, Chicago based All script analytics said "AI is the future of health care". Further she explained that how critical it will be in the field of health care management in coming years, due to the impact of various external factors, such as pollution, stressed etc. Keeping patient records only will be a too tedious task for the management. Artificial intelligence will vastly affect the hospital management and the doctors, as it will play a lead role in clinical decision making, as said by Paruk.

Machine learning, a subset of AI and data science has always been a boon to mankind and also to the healthcare industry. From saving patient data to predicting epidemic and also to propose effective medication to the patients, machine learning has always played a pivotal role when it came to the question of saving people's lives. Some of the popular applications of machine learning in the healthcare field are identifying and diagnosing diseases, drug discovery and production, smart medical record, personalised medicine, medical imaging etc[18]. Here we have used machine learning algorithms for the purpose of prediction of fatality ratio.

In this article, we have proposed a model based on machine learning algorithms to predict the mortality rate among various age groups, thus trying to help the hospital system to admit and provide essential medical facilities to people, when there are more patients than the available hospital beds. The hospital management will be able to provide intensive care to the patients and avoid critical conditions [4].Further in this paper, we will study the introduction to different models and architectures used to deduct the result and also detailed information about different classifiers used. Also, we will study the pre-processing of data including feature selection and cleaning of the data.

2. METHODS

2.1 Dataset Selection:

We have used dataset of 27891 laboratory confirmed cases of covid-19 in India of both genders and having a median age of 48.5. The disease was confirmed by real-time RT-PCR testing of virus nucleic acid. It is a nuclear derived method to detect the presence of any genetic material in a biological sample or pathogen, including a virus. The original dataset contained 12 attributes of each patient having physiological and demographic detail. At the data cleaning stage, we removed some instances which were having null values or missing values. Then the missing values were treated by Remove with Values filter in WEKA tool. After cleaning the data and treating the missing values we got 2268 instances which were useful for our model. After analyzing the data, we found that out of 2268 confirmed patients, 102 patients recovered, 44 died and 2122 were still hospitalised without any outcome.

2.2 DATA PRE-PROCESSING:

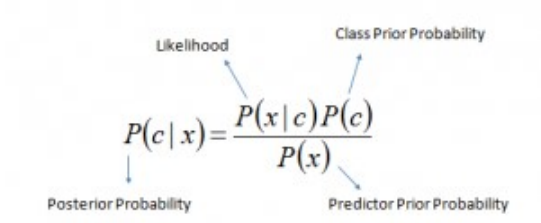
After the dataset collection, we pre-processed our model through various filter methods and classifiers. Filter methods are mostly used in the data pre-processing step. After the initial cleaning of the data, there still remain some instances and attributes which are not so useful to model. To reduce the complexity of the model, it is important to eliminate the not so relevant data. Filter methods are very popular in machine learning especially for big datasets as they are faster and more reliable than other methods and also less computationally intensive than wrapper method. The most useful parameter includes co-relation co-efficient, entropy, consistency and chi-square method [4]. Generally, in filter method the feature selection is independent of any machine learning algorithm. Instead, the features are selected through the scores obtained by the features through

different statistical parameters for their correlation with the outcome variable [5]. Here in our model after pre-processing the data some of the instances having missing values were automatically eliminated including some attributes.

2.3 CLASSIFIERS:

The Classifiersare the machine learning algorithms that are used to classify or differentiates objects into different categories based on certain features. In machine learning, classification is a supervised learning process, in which the machine learns to classify the input data based on certain parameters which are already being input in machine as test data. The input data may be of simple bi- class nature or it may be complex as multi-class nature. There are various classifiers available in machine learning. Such as linear classifiers like Linear Regression and Naive Bayes classifiers, K-Nearest Neighbour, Random Forest, J48, Support Vector Machine(SVM), Decision Tree, and Neural Networks [6]. In our model, we have used Naive Bayes classifier for pre-processing our data.

The Naive Bayes classifier is a probabilistic classification technique which is based on Bayes theorem and with an assumption among predictors. A Naive Bayes classifier assumes that the attributes are independent of all other attributes present in that model. It is not a single classifier, instead, it is a family of classifiers [7][8]. This particular family of classifiers is very scalable and particularly useful for a large data set. Mathematically it is represented as [8][10]



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

In the above equation the terms:

- P(c|x) represents the posterior probability of class (c, target) and given predictor (x, attributes).
- P(c) represents the prior probability of class.
- P(x|c) represents the likelihood which is the probability of predictor given class.
- P(x) represents the prior probability of the predictor.[10]

After cleaning and pre-processing the data we got 2268 instances with 12 unique attributes containing the demographic and physiological details of the patients. After applying the filter method in WEKA tool, some values like patient travel history, detected state, etc were removed which were not so useful for our model. After applying the filter, we got 10 unique attributes to use for our prediction system

3. PROPOSED MODEL AND IMPLEMENTATION

To have an accurate and unbiased model, we needed a balanced dataset. So, we used two different data sets for both recovered and deceased patients to train and test our model. Below figure shows the architecture of our model.

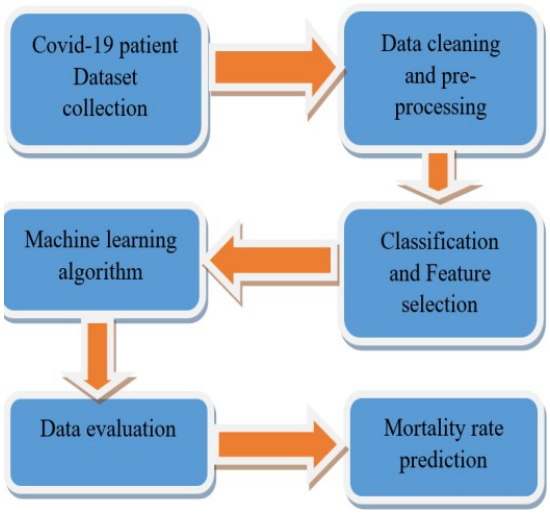


Figure 4 Model architecture

3.1. ATTRIBUTE SELECTION:

Attribute selection isalso known as Feature Selection and is the process of selecting the most appropriate attributes to be used in a model. The primary purpose of feature selection is to identify the best and most relevant features that can be added to our model and also to remove the less relevant data to make the model less complex and more accurateIn this process the classifier shows us the best valued and most relevant attributes, which can be useful for our model, simultaneously eliminating the not so relevant attributes. In this model we used Naive Bayes classifier along with Ranker Search method to select the most valued attributes.

Ranker Search Method:It is a process of feature selection in machine learning, where the features or attributes are ranked or sorted according to their effectiveness in predicting the outcome. It helps us to choose the few top attributes and discard the rest so that the model can be an effective and less complex model. After the ranker search method we found top four attributes that were useful to our model. We ran the model through various other classifiers such as Naive Bayes multinomial, J48, Random forest and found that Naive Bayes is the most useful one having an accuracy rate of 93.83%. Here is a table of accuracy rates different classifiers.

Table 1: Accuracy of different classifiers

SL NO	CLASSIFIERS	ACCURACY RATE
1	**ZeroR using 10-fold cross validation**	**93.56%**
2	**Naive Bayes using 10-fold cross validation**	**93.83%**
3	**J48 using 10-fold cross validation**	**93.56%**
4	**Random Forest using 10-fold cross validation**	**93.47%**
5	**Random Tree using 10-fold cross validation**	**93.30%**

After feature selection through the Naive Bayes classifier and Ranker search method we selected the top and most useful four attributes which included the week wise data of covid-19 cases including the age group, gender and current status of the patient such as hospitalised, recovered and deceased. Below figure shows the data processing of our feature selection.

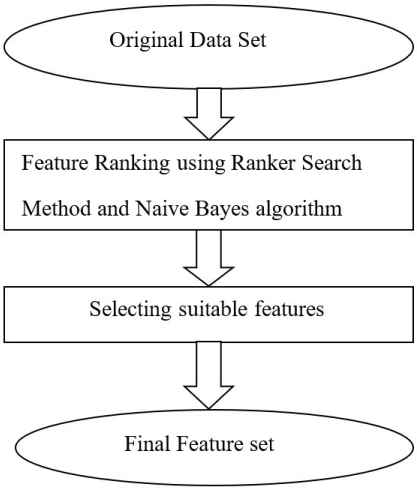


Figure 5: FEATURE SELECTION

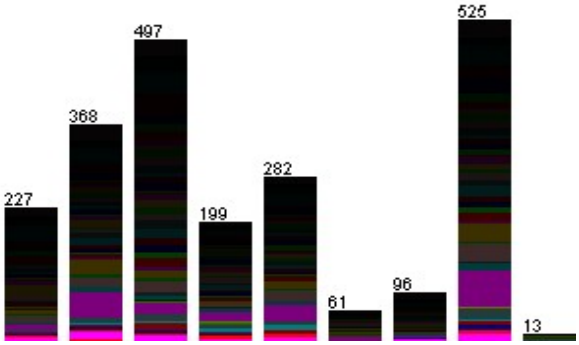


Figure 6: Graph showing death rate among covid-19 patients week wise

Above figure shows the graph of covid-19 cases among various age groups of different states, gender and nationality. It shows the number of cases rising as the weeks passed from 1st to 9th. It can be clearly seen from the above picture that how this disease is spreading rapidly over passing weeks.

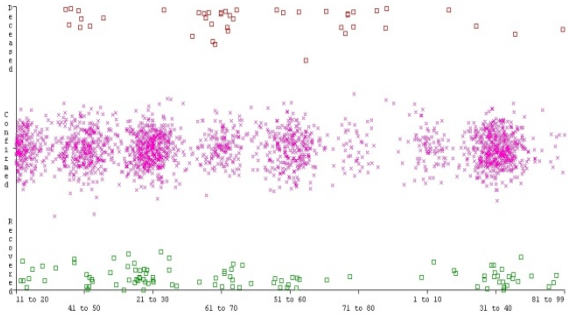


Figure 7: Plot matrix showing death rate in various age groups

This figure shows the plot matrix obtained for current status such as confirmed, recovered and deceased among various age groups over weeks. Figure 5 shows the graph of recovered and deceased patients of different age group having a median age of 48.5, minimum 1 and maximum96.

3.2.PREDICTIVE ANALYTICS:

After selecting the best and most relevant feature sets, we used various machine learning algorithms to make an accurate predictive model for predicting mortality rate among various age groups. In this model we used machine learning algorithms such as J48, Random Forest, Naive Bayes and ZeroR.

Among all classifiers used here, Naive Bayes had the most accuracy rate. We used Ranker Search method to sort and rank the best features for the model. The best Naive Bayes classification result was obtained with 2128 correctly classified instances having an accuracy rate of 93.8272 %, Mean absolute error as 0.0701 and Root mean squared error of 0.1873.Among all methods, Ranker Search method with Naive Bayes classifier has shown the most accurate result. The detailed accuracy by class is shown below.

Table 2: Detailed accuracy obtained by class

TP RATE	FP RATE	PRECISION	RECALL	F-MEASURE	MCC	ROC AREA	PRC AREA	CLASS
0.108	0.002	0.688	0.108	0.186	0.261	0.899	0.262	Recovered
0.998	0.918	0.940	0.998	0.968	0.227	0.860	0.998	Confirmed
0.000	0.000	0.000	0.000	0.000	-0.003	0.825	0.094	Deceased
Weighted Average	0.938	0.859	0.911	0.938	0.914	0.224	0.862	0.938

The model was evaluated using 10-fold random cross validation (with no replacement and no overlap). We calculated the overall accuracy of all classifiers of machine learning used here. We also generated the Receiver Operating Characteristic (ROC) curve and confusion matrix to evaluate the accuracy of our model and. We made sure that there are no duplicate values in the dataset.

4. RESULTS AND ANALYSIS:

The purpose of the model is to develop a predictive system to forecast the mortality rate among the covid-19 patients in India of various age groups. The aim was to help the hospital management to provide intensive medical care to the more vulnerable patients and tackle the pandemic effectively. As discussed earlier several measures like Accuracy, Confusion Matrix, AUC and ROC are used to test the model. In table 3 the confusion matrix showing the correctly classified instances and classes are shown below:

Table 3: Confusion matrix of Naive Bayes classifier

a	b	c	<-classified as
11	91	0	a=Recovered
4	2117	1	b= Confirmed
1	43	0	c= Deceased

A confusion matrix is a table that used for summarization of the performance of a classifier. A Confusion Matrix shows the ways in which the classification gets confused while making prediction. Below figure shows the plot area under ROC of false positive rate vs true positive rate of class value deceased. A receiver operating characteristic or ROC curve is a graphical representation of the performance of a binary classification system.

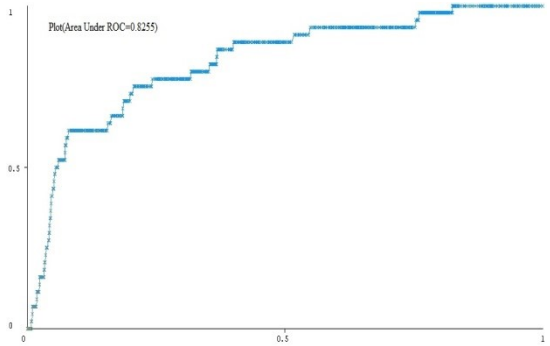


Figure 8: ROC curve showing true rate vs false rate of class deceased

The result demonstrates that the above model is able to serve its purpose of predicting the mortality rate among COVID-19 patients of India basing upon the age group.

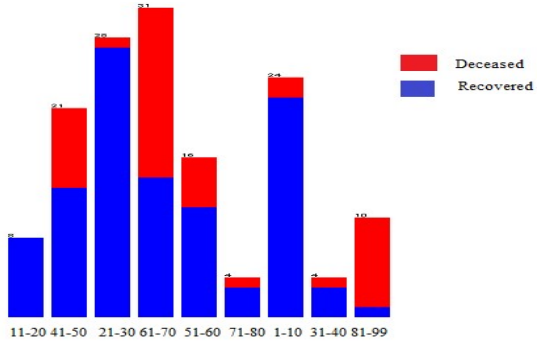


Figure 9: Graph showing mortality rate of class age

5. CONCLUSION AND FUTURE WORK

After going through the model, we found that patients who come under the older age group such as patients in their 40's, 50's, 60's and 80's have high mortality rate. Especially we can say the patients who are senior citizens or above 60's has high chances of facing the fatality.

The model can be developed further for predicting the mortality rate among patients having pre-existing and mortality rate among people of different provinces. As per wordometers information, death rate from COVID-19 in India is still as lower as 9% and the recovery rate is 91% [9]. The reason behind the lower death rate can be the hot and humid climate of India, which is beneficial for the country in some manner. Still to avoid crowded situation and to save lives of people we need to provide ultimate medical care to the patients who belong the high mortality group. This model can be improved more to accommodate more datasets of patients from across the globe and having different types of pre-existing medical conditions.

We can further develop it to predict the death rates among various other medical conditions such as diabetes, heart disease, lungs weakness and also with cancer. In a densely populate country like India where the medical facility is not easily available for people on time and where thousands of people die due to lack of proper medical care, this model can help the hospital and medical staff to effectively take a decision to hospitalise patients in priority basis, so that the precious lives can be saved and also the mortality rate curve can be flattened with the mixed effort of doctors and the scientists.

REFERENCES

[1] https://en.wikipedia.org/wiki/Coronavirus

[2] https://en.wikipedia.org/wiki/2019%E2%80%9320_coronavirus_pandemic]

[3] https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it

[4] Pourhomayoun, Mohammad, and Mahdi Shakibi. "Predicting Mortality Risk in Patients with COVID-19 Using Artificial Intelligence to Help Medical Decision-Making." medRxiv (2020).

[5] https://www.analyticsvidhya.com/blog/2016/12/introduction-to-feature-selection-methods-with-an-example-or-how-to-select-the-right-variables/

[6] https://medium.com/@Mandysidana/machine-learning-types-of-classification-9497bd4f2e14

[7] https://en.wikipedia.org/wiki/Naive_Bayes_classifier

[8] https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/

[9] https://www.worldometers.info/coronavirus/country/india/

[10] www.tandfonline.com

[11] affectivetweets.cms.waikato.ac.nz

[12] www.researchsquare.com

[13] www.robesonian.com

[14] en.wikipedia.org

[15] https://en.wikipedia.org/wiki/COVID19_pandemic_in_India#Total_confirmed_cases,_active_cases,_recoveries_and_deaths

[16] https://theprint.in/opinion/current-rate-india-30000-covid-19-deaths-may-no-hospital-bed-june-data/385386/

[17] https://en.wikipedia.org/wiki/COVID-19_pandemic_in_India#/media/File:India_COVID-19_cases_density_map.svg

[18] https://builtin.com/artificial-intelligence/machine-learning-healthcare