Prediction and Diagnosis of COVID-19 using Machine Learning Algorithms

Heet Savla, Vruddhi Mehta, Ramchandra Mangrulkar

Abstract**:** ***The world is reworking in a digital era. However, the field of medicine was quite repulsive to technology. Recently, the advent of newer technologies like machine learning has catalyzed its adoption into healthcare. The blending of technology and medicine is facilitating a wealth of innovation that continues to improve lives. With the realm of possibility, machine learning is discovering various trends in a dataset and it is globally practiced in various medical conditions to predict the results, diagnose, analyze, treat, and recover. Machine Learning is aiding a lot to fight the battle against Covid-19. For instance, a face scanner that uses ML is used to detect whether a person has a fever or not. Similarly, the data from wearable technology like Apple Watch and Fitbit can be used to detect the changes in resting heart rate patterns which help in detecting coronavirus. According to a study by the Hindustan Times, the number of cases is rapidly increasing. Careful risk assessments should identify hotspots and clusters, and continued efforts should be made to further strengthen capacities to respond, especially at sub-national levels. The core public health measures for the Covid-19 response remain, rapidly detect, test, isolate, treat, and trace all contacts. The work presented in this paper represents the system that predicts the number of coronavirus cases in the upcoming days as well as the possibility of the infection in a particular person based on the symptoms. The work focuses on Linear Regression and SVM models for predicting the curve of active cases. SVM is least affected by noisy data, and it is not prone to overfitting. To diagnose a person our application has a certain question that needs to be answered. Based on this, the KNN model provides the maximum likelihood result of a person being infected or not. Tracking and monitoring in the course of such pandemic help us to be prepared.***

Keywords: Healthcare, K-Nearest Neighbor, Linear Regression, Machine Learning Support Vector Machine.

I. INTRODUCTION

The ability of Science and Technology to improve human life is known to everyone and hence the use of technologies is increasing day by day. Machine Learning is one such field of technology that has become popular in a very short period of time. Machine learning is the process of teaching machines to identify patterns by providing them data and an algorithm to work with the data. It is a tool that is used to transform information into knowledge. Machine learning is now used in almost every sector of life including education, finance, transportation, etc. Lately, machine learning techniques have also been adopted by the Healthcare sector.

Revised Manuscript received on August 01, 2020.
Revised Manuscript received on August 05, 2020.
Manuscript published on September 30, 2020.
* Correspondence Author

Heet Savla*, Department of Computer Engineering, K.J. Somaiya College of Engineering, Vidyavihar, Mumbai, Maharashtra, India, Email: heetsavla99@gmail.com

Vruddhi Mehta, Department of Computer Engineering, Dwarkadas J. Sanghvi College of Engineering, Mumbai, Maharashtra, India, Email : vruddhimehta12@gmail.com

Ramchandra Mangrulkar, Department of Computer Engineering, Dwarkadas J. Sanghvi College of Engineering, Mumbai, Maharashtra, India, Email: ramchandra.mangrulkar@djsce.ac.in

The fusion of Machine learning with healthcare has resulted in a great outburst of applications that should lead to medical revolutions. There are abundant applications of Machine learning in healthcare such as Identification of Diseases and Diagnosis, Medical Imaging, Drug Discovery and Manufacturing, personalized medicine treatment, smart health records, etc. [3]

Viral pandemics are a serious threat. COVID-19 is not the first, and it won't be the last. Hundreds of research teams around the world are combining their efforts to collect data and develop solutions. Machine learning is helping to fight against this pandemic in various ways which are :

A. Identify the risk

Machine learning has been proven very valuable in predicting risks in various realms. With medical risk precisely, machine learning is potentially interesting in three key ways.

- The virus risk: This is used to find the risk in a specific individual or group infected with coronavirus.
- The severe risk: This is used to find the risk in a specific individual or a group which are developing extreme COVID-19 symptoms.
- Outcome risk: This is used to check how an individual or a group will respond to a specific treatment.

B. Diagnosticate patients

Machine learning is used to diagnose patients in different ways like:

- Using temperature scanners for identifying if a person has fever,
- Using wearable technology connected with smart applications to monitor patient’s heart rate at rest,
- Using machine learning powered chatbots to generate a self report by asking a few questions based on symptoms experienced.

C. Developing drug faster

Machine learning will speed up the drug development process considerably while not sacrificing elementary control.

Earlier, during the spread of Ebola virus, researchers were trying to discover small molecule inhibitors. They observed that training Bayesian Learning models with viral pseudotype entry assay and the Ebola virus replication assay data helped speed up the scoring process.

The model quickly identified three potential molecules for testing. During such times of pandemic, getting more accurate scores faster is critical to speeding up drug development.

D. Finding existing drugs that can help.

There is a lot of time and money spent by the companies to create the drug and get it approved. This is because the companies have to be very sure that the drug won’t have any harmful side effects. However in a situation like pandemic, a faster response is needed and developing a new drug in such a short span is not possible. As an alternative, the drugs which are already used to treat other similar diseases are used. However, there are tons of drugs available and to find which drug can be effective, ML is used. ML can aid in prioritizing drug candidates much faster by automatically:

- Constructing a knowledge graph which helps scientists to identify the connection between the virus and a drug candidate.
- By identifying drug-target interactions (DTIs) between the virus’s proteins and existing drugs to predict possible drug candidates for a new vaccine.

E. Predicting the spread of virus

Social media analysis can be done to assess the likelihood of novel virus contamination. Content can be interpreted from the social communications. The machine learning model might not be able to categorize people on an individual level, but it can use all of this data to estimate the spread of the pandemic in realtime and to project the ecalation in the forthcoming weeks.[3]

II. LITERATURE SURVEY

Ramesh et al. proposed an ensemble method based predictive model for analyzing disease datasets. The authors use various machine learning approaches to predict the disease. Their work proposes a predictive model using the preprocessing techniques. The performance of the model is further improved by using ensemble methods. The authors have used 10 fold validation technique. The classifications algorithms used are SVM, KNN, Naïve Bayes, Decision Tree, and Random Forest. They perform a predictive analysis on the datasets such as Diabetes, ILPD (Indian Liver Patient Disease), CKD (Chronic Kidney Disease), Hepatitis disease, Cardiovascular Disease (CVD) or heart, Cancer disease. Their experimental results show that the proposed predictive model outperforms in terms of better accuracy.[4]

F. Rustam, et al. presented the work based on COVID-19 Future Forecasting Using Supervised Machine Learning Algorithms. The authors have made three predictions. They have predicted the number of infected cases, deaths, and recoveries. The four methodologies used for comparison are linear regression (LR), least absolute shrinkage and selection operator (LASSO), support vector machine (SVM), and exponential smoothing (ES). The author exhibits that these prediction systems can be very useful in decision making to supervise the present scenario to guide early interventions to manage these diseases very effectively. The authors go on further explaining the techniques. Linear Regression provides a linear relationship between the dependent and independent variables. These two factors are mainly involved in it. They also state that to get the best fit implies that the difference between the actual values and predicted values should be minimum. The LASSO approach shrinks the absolute values of a data specimen to central values. The features not helping in the regression are made potentially equal to zero. SVM applies not only to regression but also to classification. The set of functions called kernel transforms the data inputs into the desired form. Lastly, exponential smoothing forecasting is done based on previous periods of data. The authors feel it is a very efficient algorithm for univariate data. The author evaluates the performance of each of the learning models in terms of R-squared (R2) score, Adjusted R-Square (R2adjusted), mean square error (MSE), mean absolute error (MAE), and root mean square error (RMSE). After several implementations, the author proves that ES performs best in the current forecasting domain, given the nature and size of the dataset.[6] Ahmad S et al. has proposed a simple algorithm helps early identification of SARS-CoV-2 infection patients with severe progression tendency. The authors have designed a simple algorithm that facilitates early identification of COVID-19 progression and aimed to adjust the huge glide of infected patients between primary health care and tertiary centers. The authors have gathered data from the Shanghai Public Health Clinical Center. The dataset consisted of blood count, lymphocyte subsets, C-reactive protein (CRP), procalcitonin (PCT), alanine aminotransferase (ALT), aspartate aminotransferase (AST), gamma glutamyl-transpeptidase (GGT), lactate dehydrogenase (LDH), total bilirubin (TBIL), creatinine, creatine kinase (CK), and D-dimer was obtained with data collection forms. The authors observed that severe cases dealt with respiratory distress, pulse oxygen saturation, oxygenation index, require mechanical ventilation, and shock. The authors perform statistical analysis on it. The authors proposed a model based on three routine parameters: age, LDH, and CD4 count. The authors found that about 50% of patients with age-LDH-CD4 model ≥ 82 will progress to severe cases and can benefit from early transfer to tertiary centers. The authors showed that SARS-CoV-2 infection had a low severe rate and fatality rate once the control measures (early discovery, early reporting, early quarantine, and early treatment) were undertaken at the beginning of COVID-19 outbreaks. The age, LDH, and CD4 counts were the independent predictors of COVID-19 progression. The authors conclude stating that the algorithm is a simple and accurate index for the early diagnosis of patients. [7]

III. PROPOSED METHODOLOGY

A. Predicting the COVID-19 Cases

This project considers two methodologies : Linear Regression and SVM model. SVM produces more accurate results because it is least affected by noisy data, and it is not prone to over fitting. The support vector regression can be set apart from SVM with only a few deviations. Firstly, the result is a real number.

Retrieval Number: 100.1/ijrte.C4676099320
DOI:10.35940/ijrte.C4676.099320
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication

Therefore, it becomes very challenging to predict the information at hand. For the case of regression, set an approximation of the epsilon. Besides this fact, the algorithm is too complicated to be taken into consideration.

However, the main idea always remains the same that is to minimize errors and maximize the margin. This can be done by individualizing the hyper planes, keeping in mind that part of the error is tolerated. [5]

SVM uses different parameters to build the model such as kernel, c, gamma, epsilon, shrinking, etc. The parameter C is for controlling the outliers. Lower values of C indicate that it is allowing more outliers. Higher values indicate it is allowing only some outliers. For implementation C is taken as 1. Next, the polynomial kernel function is used with degree 6. The degree of polynomial determines flexibility. The value of epsilon is directly proportional to errors. Epsilon is considered as 0.01 for our case, which helps us minimize the error and maximize the margin. The other parameters are default.

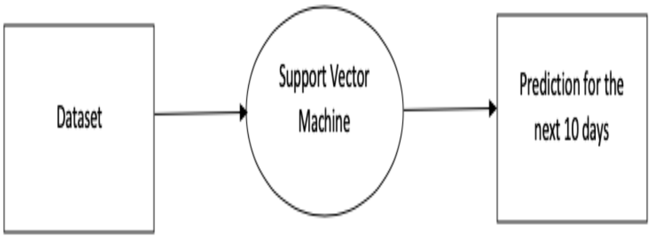


Fig. 1. Block Diagram to predict confirmed cases using SVM

Linear Regression is a type of supervised machine learning algorithm. It is used to forecast based on cause and effect relationship between variables. It makes a linear line with a constant slope. The data points are used for further prediction. The number of cases is predicted based on the slope of the line. Normalization is set to true. The regressors will be normalized by subtracting the mean and dividing by the l2-norm. By using machine learning you can visualize the data and graphs for many things but here the disadvantage is the high error susceptibility which leads to giving an error in the predictions. [9]

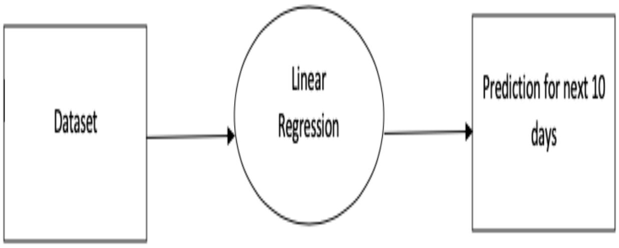


Fig. 2. Block Diagram to predict confirmed cases using LR

B. Diagnosis of COVID-19

KNN is the most fundamental type of case-based learning, instance-based learning, or lazy learning. The KNN technique is applied for classifying patients based on the closest training dataset in the feature space. It assumes all cases are points in n-dimensional space. A distance measure K is needed to determine the "closeness" of instances. KNN classifies a new patient by finding its nearest neighbors and picking the most popular class among the neighbors.

A random instance is denoted by(p1(x), p2(x), p3(x),.., pn(x)), where pi(x) denotes features. Our dataset has five features: fever, body pain, age, cough, and cold, breathing issues. The Euclidean distance is between two points X and Y given by equation d(X, Y)=∑i=1n(Xi−Qi)^2. The K nearest neighbor algorithm is the simplest of all machine learning algorithms, and it is analytically tractable. [1]

KNN is highly receptive to the local information. A new record can be estimated using the nearest neighbor, which runs parallel because each data point is independent of the other. Therefore, the algorithm can take advantage of the data provided prior and form highly adaptive and highly nonlinear decision boundaries for each data point. The algorithm checks each patient record against the testing dataset for the nearest neighbor. The manipulation of the value of K makes a huge difference. The right chosen value gives the maximum accuracy. Here, the value of K is taken as 8, which gives an accuracy of 54%. Further, experimentations shall provide with the higher accuracy score. [2]

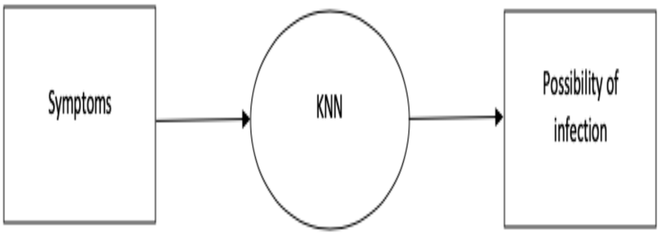


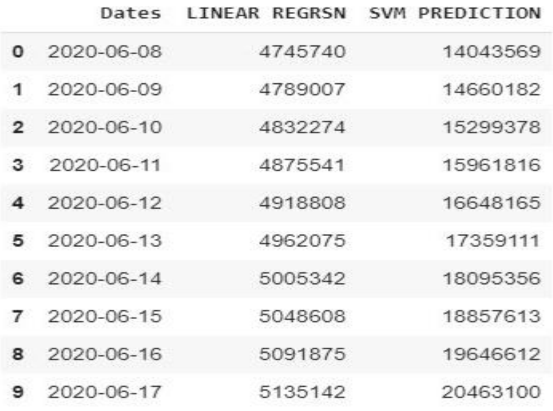
Fig. 3. Block Diagram to diagnose people

IV. EXPERIMENTAL SETUP AND RESULTS

A. Predicting the COVID-19 Cases

Out of all precautions, "be informed" about all the aspects of COVID-19 is considered extremely important. This study attempts to develop a system for the future forecasting of the number of cases affected by COVID-19 using machine learning methods. The dataset used for the study contains information about the daily reports of the number of newly infected cases, the number of recoveries, and the number of deaths due to COVID-19 worldwide. As the death rate and confirmed cases are increasing day by day which is an alarming situation for the world. The number of people who can be affected by the COVID-19 pandemic in different countries of the world is not well known. John Hopkins University has made the data available for educational and academic research purposes. The main file is covid_19_data.csv. This study is an attempt to forecast the number of people that can be affected in terms of new infected cases and deaths including the number of expected recoveries for the upcoming 10 days. We have implemented in the Google Colab environment. Two machine learning models LR and SVM have been used to predict the number of newly infected cases, the number of deaths, and the number of recoveries [10]. Accuracy, F1 Score, Recall, and Precision, etc. Are a number of the metrics used to evaluate 'Classification Machine Learning Problems' in which the output/target column is categorical. The problem which is solved here is a classic 'Regression Machine Learning Problem' in which the output/goal column is continuous. For these varieties of problems metrics for evaluation is Root Mean Square Error.

The implementation has expected the range of cases for the following few days. This allows us to prepare for future arrangements to be done. The table given under depicts the predictions.



	Dates	LINEAR REGRSN	SVM PREDICTION
0	2020-06-08	4745740	14043569
1	2020-06-09	4789007	14660182
2	2020-06-10	4832274	15299378
3	2020-06-11	4875541	15961816
4	2020-06-12	4918808	16648165
5	2020-06-13	4962075	17359111
6	2020-06-14	5005342	18095356
7	2020-06-15	5048608	18857613
8	2020-06-16	5091875	19646612
9	2020-06-17	5135142	20463100

Fig. 4. Output of prediction for the next 10 days.

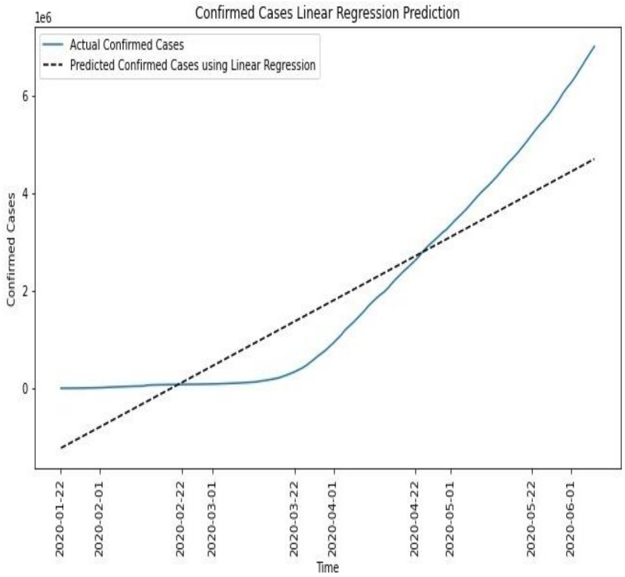


Fig. 5. Prediction curve for LR

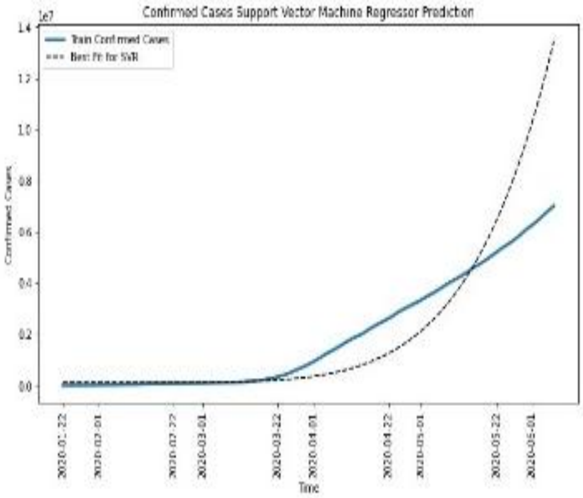


Fig. 6. Prediction curve for SVM

SVM is a useful and flexible technique, helping the user to deal with the limitations of distributional properties of underlying variables, the geometry of the data, and the common problem of model over fitting. The selection of kernel functions is significant for SVR modeling. It is observed that SVM is superior to LR as a prediction method. LR cannot capture the nonlinearity in a dataset and SVM becomes handy in such situations. Compute Root Mean Square Error (RMSE) for both LR and SVM models to evaluate the performance of the models. The values obtained are shown in the figure given below.

Root Mean Square Error for Linear Regression: 1830971.42586053

Root Mean Square Error for Support Vector Machine: 4225835.260279156

B. Diagnosis of COVID-19

It is very crucial to identify cases on an early basis. According to the circumstances, people fear getting tested. They try all the homecare until the last moment. So when they are brought to the hospital they need the utmost attention. This is causing panic and fewer number of ICU's available. If a person is warned using the application, then he can be prepared mentally and get testing done at an early stage. The usage of this application can encourage people to get tested early. The recovery rate will increase if people are aware of it. The people going to the hospital may not need intensive care. There will be no crunch in the availability of ICU's. There will be less panic and a healthier environment. The dataset is obtained from Kaggle. The dataset contains five major variables that will be having an impact on whether someone has coronavirus disease or not. According to WHO, 5 are major symptoms of COVID-19, Fever, Tiredness, Difficulty in breathing, Dry cough, and Age. 2000 patient records have been taken for our application. The environment used for execution is Google Colab. The user who checks for his symptoms is matched with the dataset for the nearest person with such symptoms and predicts the output. The results are categorized into two, possible, or not possible. The table given below represents one of our testing data.

Parameters	Fever	Body Pain	Age	Cough and Cold	Breathing issues	Possibility
Person 1	104.3	1	78	0	0	60.6%
Person 2	90.1	1	65	1	-1	44.5%
Person 3	98.5	1	22	0	-1	50.6%
Person 4	112.5	1	86	1	1	65%
Person 5	96	1	35	0	-1	48%
Person 6	108.3	0	18	0	0	54%
Person 7	89.8	1	85	1	-1	45.5%

Fig. 7. Input Table

Retrieval Number: 100.1/ijrte.C4676099320
DOI:10.35940/ijrte.C4676.099320
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication

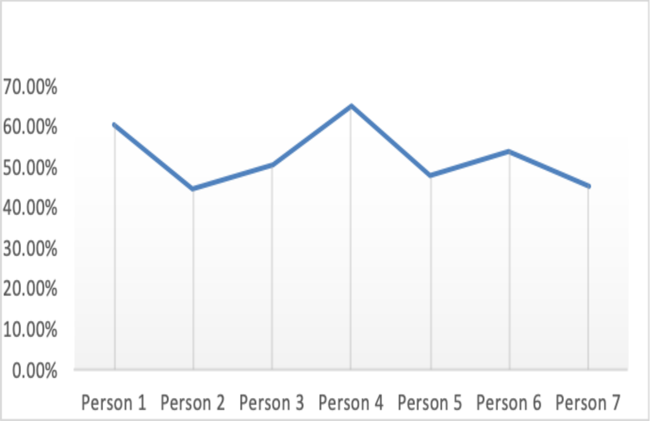
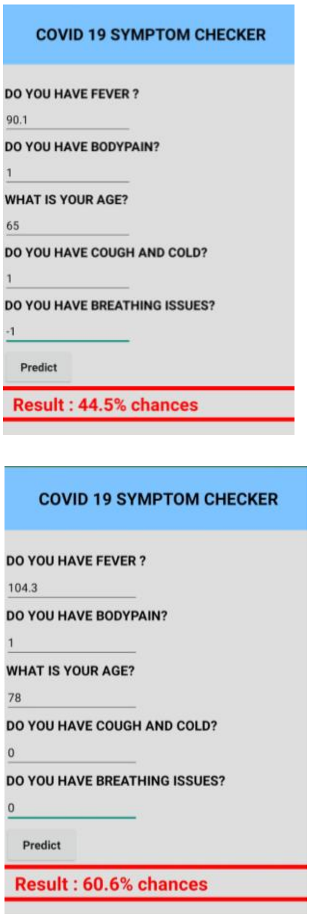


Fig.8. Output graph for 5 people.



COVID 19 SYMPTOM CHECKER

DO YOU HAVE FEVER ?
90.1
DO YOU HAVE BODYPAIN?
1
WHAT IS YOUR AGE?
65
DO YOU HAVE COUGH AND COLD?
1
DO YOU HAVE BREATHING ISSUES?
-1

Result : 44.5% chances

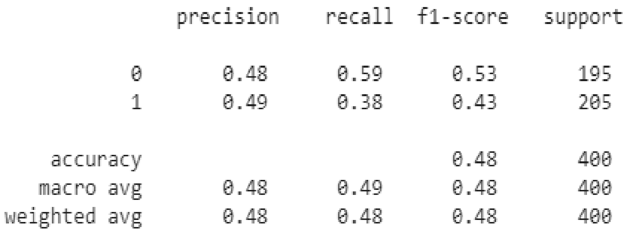
COVID 19 SYMPTOM CHECKER

DO YOU HAVE FEVER ?
104.3
DO YOU HAVE BODYPAIN?
1
WHAT IS YOUR AGE?
78
DO YOU HAVE COUGH AND COLD?
0
DO YOU HAVE BREATHING ISSUES?
0

Result : 60.6% chances

Fig. 9. Snapshot of the android application

The accuracy of our model can be found using the confusion matrix. The various other parameters like precision, recall, f1 score and support are also calculated. The given figure is an evaluation of our mode. The model has an accuracy of 48%. It can be increased by changing the value of k.

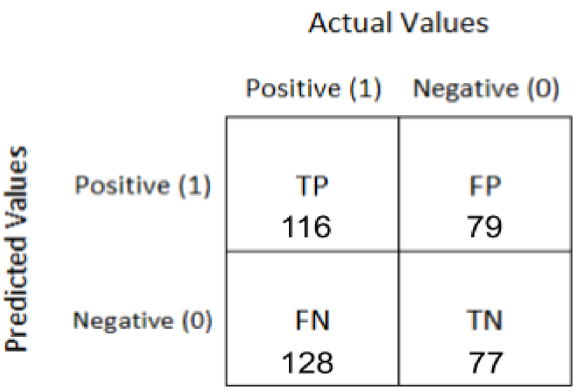


```
              precision    recall  f1-score   support

           0       0.48      0.59      0.53       195
           1       0.49      0.38      0.43       205

    accuracy                           0.48       400
   macro avg       0.48      0.49      0.48       400
weighted avg       0.48      0.48      0.48       400
```

Fig. 10. Snapshot of evaluation of model



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP 116	FP 79
	Negative (0)	FN 128	TN 77

Fig. 11. Confusion Matrix for KNN model

Results are obtained based on 5 parameters only. In the future, running on the clinical data will provide maximum possible efficiency. A larger dataset provides a better fit for the algorithm.

V. CONCLUSION

In this paper, the analysis was performed based on the data set provided by the Johns Hopkins Corona Virus Resource Center. Some of the most popular machine learning algorithms were deployed and were able to achieve considerably good results. After reviewing many research papers and implementing the system, concluding that this prediction can help us take much care against the virus as per the prediction results. The diagnosing system can be very useful in the initial stage of the virus and help us know the possibility of infection. The challenge in the analysis of this data set is that it is growing by the day and the number of cases is increasing exponentially. Several future works include the implementation of the system using other machine learning models that gives much more precision. The scope extends further to bringing this application into the market for the people to self-diagnose themselves.

REFERENCES

1. Krati Saxena, Dr. Zubair Khan, Shefali Singh," Diagnosis of Diabetes Mellitus using K Nearest Neighbor Algorithm," International Journal of Computer Science Trends and Technology (IJCST) – Volume 2 Issue 4, July-Aug 2014, Page 36, ISSN: 2347-8578.
2. M. Nirmala Devi, S. A. alias Balamurugan and U. V. Swathi, "An amalgam KNN to predict diabetes mellitus," 2013 IEEE International Conference ON Emerging Trends in Computing, Communication and Nanotechnology (ICECCN), Tirunelveli, 2013, pp. 691-695, DOI: 10.1109/ICE-CCN.2013.6528591.
3. V. Chamola, V. Hassija, V. Gupta and M. Guizani, "A Comprehensive Review of the COVID-19 Pandemic and the Role of IoT, Drones, AI, Blockchain, and 5G in Managing its Impact," in IEEE Access, vol. 8, pp. 90225-90265, 2020, doi: 10.1109/ACCESS.2020.2992341.
4. Ramesh, D., Katheria, Y.S. Ensemble method based predictive model for analyzing disease datasets: a predictive analysis approach. Health Techno. 9, 533–545(2019).https://doi.org/10.1007/s12553-019-00299-3
5. S. S. Arun and G. Neelakanta Iyer, "On the Analysis of COVID19 - Novel Corona Viral Disease Pandemic Spread Data Using Machine Learning Techniques," 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2020, pp. 1222-1227, doi: 10.1109/ICICCS48265.2020.9121027.

6. F. Rustam et al., "COVID-19 Future Forecasting Using Supervised Machine Learning Models," in IEEE Access, vol. 8, pp. 101489-101499, 2020, DOI: 10.1109/ACCESS.2020.2997311.
7. Ahmad S, Hafeez A, Siddqui SA, Ahmad M, Mishra S. A Review of COVID-19 (Coronavirus Disease-2019) Diagnosis, Treatments and Prevention. EJMO 2020;4(2):116–125.
8. M.K., Arti. (2020). Modeling and Predictions for COVID 19 Spread in India. 10.13140/RG.2.2.11427.81444.
9. Ghosal S., Sengupta S., Majumder M., Sinha B. Linear Regression Analysis to predict the number of deaths in India due to SARS-CoV-2 at 6 weeks from day 0 (100 cases - March 14th 2020) Diabetes Metab Syndr. 2020;14:311–315. doi: 10.1016/j.dsx.2020.03.017.
10. Lixiang Li, Zihang Yang, Zhongkai Dang, Cui Meng, Jingze Huang, Haotian Meng, Deyu Wang, Guanhua Chen Jiaxuan Zhang, Haipeng Peng, Yiming Shao, Propagation Analysis and Prediction of the COVID-19, Infectious Disease Modelling, Volume 5, 2020, Pages 282-292.

AUTHORS PROFILE

Mr. Heet Savla, has gained his diploma in Computer Engineering. He is currently in her 4 th year of computer engineering. He is pursuing B. TECH degree from K.J. Somaiya College of Engineering, Vidyavihar, Mumbai, India. He has done a lot of research projects in analytics, machine learning and software development arena. He has presented papers at various conferences. The topic of his research paper includes image search, IoT, and face detection. Data analytics, business analytics and machine learning are her main area of interests.

Ms. Vruddhi Mehta, has completed her Diploma in Computer Engineering from Shri Bhagubhai Mafatlal Polytechnic, Mumbai, Maharashtra, with distinction. She is pursuing a BE in Computer Engineering from Dwarkadas J. Sanghavi College of Engineering, Mumbai, Maharashtra. She has published various papers in the fields of blockchain, databases, and cybersecurity in reputed journals. Her area of interest lies in artificial intelligence and cybersecurity.

Dr. Ramchandra Mangrulkar, being a post graduate from National Institute of Technology, Rourkela Odisha, received his PhD in Computer Science and Engineering from SGB Amravati University, Amravati in 2016 and presently he is working as an Associate Professor in the department of Computer Engineering at SVKM's Dwarkadas J. Sanghvi College of Engineering, Mumbai (Autonomous College affiliated to University of Mumbai), Maharashtra, India. Prior to this, he worked as an Associate Professor and Head, department of Computer Engineering, Bapurao Deshmukh College of Engineering Sevagram. Maharashtra, India. Dr. Ramchandra Mangrulkar has published 48 papers and 12 book chapters with Taylor and Francis, Springer and IGI Global in the field of interest. Also presented significant papers in related conferences. He has also chaired many conferences as a session chair and conducted various workshops on Artificial Intelligence BoT in Education, Network Simulator 2 and LaTeX and Overleaf. He has also received certification of appreciation from DIG Special Crime Branch Pune and Supretendant of Police. He is also ICSI-CNSS Certified Network Security Specialist. He has also received grant in aid of Rs. 3.5 laks under Research Promotion Scheme of AICTE, New Delhi for the project "Secured Energy Efficient Routing Protocol for Delay Tolerant Hybrid Network". He is active member of Board of Studies in various universities and autonomous institute in India.