COVID-19 Severity Prediction with Machine Learning.

Raghavendra A, Samarth D Valmiki, R Hemanth Kumar, Kuruva Hithendra SaiKumar and Harshini R
School of Computer Science and Engineering, REVA University

***Abstract* — Coronavirus sickness has been labelled an infectious pandemic, affecting millions of people around the world. Healthcare institutions, particularly in underdeveloped nations, are at risk of exceeding their limit and capacity due to a lack of vaccines and rapid virus transmission from person to person. As a result, it is critical to appropriately manage resources in these countries in order to limit the high death rate and the resulting damage. It has resulted in a large number of deaths, resulting in a global state of emergency. In order to deal with the limitation of resources, we took data from COVID-19 positive patients and constructed and applied a machine learning classification model to forecast the severity of the illness.**

***Index Terms* — Coronavirus, Severity, Framework, Pandemic, Predictive models.**

I. Introduction

The novel Coronavirus disease was first reported in China, in December 2019. It quickly spread over the globe. The causal virus's cumulative incidence quickly raised and affected 196 nations and territories, with the United States, Spain, Italy, United Kingdom, and France among the most affected followed by India. The World Health Organization declared the outbreak of the coronavirus a pandemic as the virus continued to spread. New solutions are needed to generate, manage, and analyze large amounts of data on the incidence of conditions, patient data, and community activity, as well as to integrate clinical trial data, medications, genetics, and public health data. Researchers can predict where and when the disease will spread by combining this data with machine learning (ML) and artificial intelligence (AI) and warning those areas to be prepared.

The main motivation for the COVID-19 Patient Severity prediction based on the patient's record is the global havoc it has created, and the suffering people are undergoing through this. In this paper, we are trying to build a Machine Learning model to predict the severity of a patient based on their previous medical records and travel history. As it is known that a patient's previous health state plays a gargantuan role in deciding whether a person can survive the pandemic, we have decided to use Machine learning to check and tell who suffers the most specific and allied factors for the severity of infection on the patient.

II. Literature Review

[1] COVID-19 Outbreak Prediction with Machine Learning provides many outbreak prediction modals for COVID-19, which may help people around the area to make mindful selections and take measures. Among the same old fashions for COVID-19 worldwide pandemic prediction, easy epidemiological and statistical fashions have obtained greater interest from authorities, and they're famous withinside the media. Due to an excessive degree of uncertainty and shortage of crucial information, popular fashions have proven low accuracy for long-time period prediction. This paper affords a comparative evaluation of gadget getting to know and smooth computing fashions to are expecting the COVID-19 outbreak as an opportunity to SIR (Susceptible, Infected, Recovered) and SEIR (Susceptible, Exposed, Infectious, Recovered) fashions [1].

[2] COVID-19 occurrence forecasting using Supervised Machine Learning Models demonstrates the functionality of Machine Learning fashions to forecast the variety of upcoming sufferers laid low with COVID-19 that's currently taken into consideration as a capability risk to mankind. Three kinds of predictions are made via way of means of each of the fashions, along with the variety of newly inflamed instances, the variety of deaths and recoveries within the subsequent 10 days [2].

[3] Prediction version and chance rankings of ICU ward admitting and mortality index in COVID-19 cope with a retrospective assessment of clinical statistics of geography, laboratory assessments and comorbidities on the preliminary presentation. The number one results had been ICU admission and loss of life. Logistic regression become used to perceive impartial scientific variables predicting the 2 results [3].

[4] Machine-learning prediction of coronavirus disease in India. Makes an easily available device which lets one know ways to expect the type, size, and timeline of COVID-19 instances

volume and wind-up length crosswise India. Method outperformed whilst likened to formerly available sensible fashions on the bases of precision of prediction. Hence, installing area the measures of prevention can successfully manipulate the unfold of COVID-19, and additionally, the loss of life price could be decreased and, in the end, be over in India and different nations [4].

[5] Modelling and Prediction of the disease with Deep Assessment Methodology and fractional calculus makes a speciality of modelling, predicting, and evaluating confirmed, recovered, and useless instances of COVID-19 via way of means of the use of Fractional Calculus in contrast with different fashions for 8 countries [5].

[6] Development and validation of a gadget getting to know-primarily based prediction version for near-time period in-health centre mortality amongst sufferers with COVID-19. A version is evolved to validate the predictions of near-time period in-health centre mortality amongst sufferers with COVID-19 via way of means of the utility of a gadget getting to know (ML) set of rules on time-series inpatient information from digital fitness statistics [6].

[7] Covid-19 of Portugal province: prediction of hospital admission, ICU and predictions on respiratory-health made on the numerous ranges of a patient's data, namely: pre-hospitalization (checking out time), post-hospitalization, and post-extensive care. The well-timed prediction of the clinical wishes of inflamed people permits a higher and faster care provision for the essential instances, helping the control of to be had resources [7].

III. Methodology

Datasets are important in the process of selecting the perfect algorithm for the problem. The Random Forest algorithm is used to classify the patient's records. Many Machine Learning techniques can be applied to this problem statement. Machine learning techniques include supervised learning, regression, etc. The appropriate technology that can be used for this problem statement is the supervised machine learning technique. To categorize the problem by input and/or output and to understand the data. Some algorithms can work with small0 sample sets while others require huge amount of samples. After exploring the dataset, the selection of features plays an important role. A computational model will be developed by applying the algorithms that will efficiently detect patient healthcare reports. Several supervised learning algorithms are proposed among which an efficient algorithm is selected which helps us to achieve better accuracy effectively.

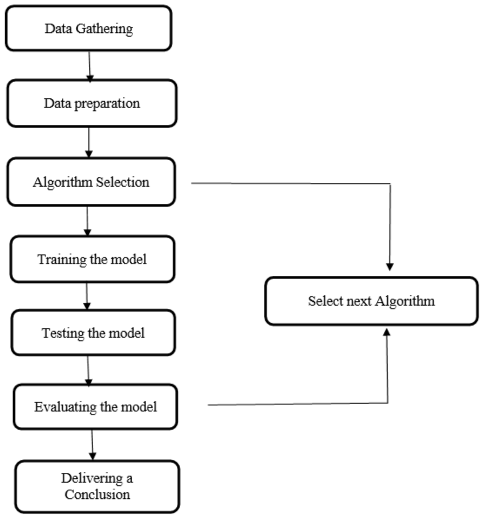


Fig 1.1 System Architecture

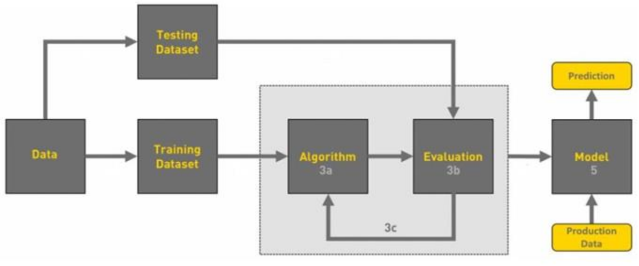


Fig 1.2. Data Flow Diagram

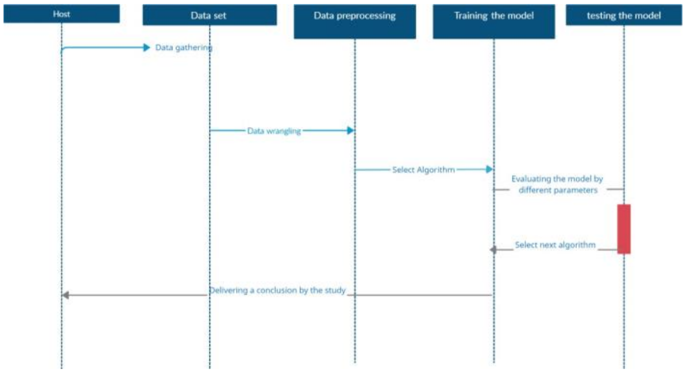


Fig 1.3. Use Case Diagram

A. *Random Forest (R.F)*

R.F are a combination of algorithms for classification, regression etc., that do the job by creating many decision trees at the time of training and giving out the class which Classification or Regression of the individual trees within.

Random forests correct from decision trees' property of overfitting to its training data set.

A supervised classification algorithm, the Random Forest algorithm is used to create a forest in some fashion and make it random, as evidenced by its name. The amount of trees in the forest has a direct relationship with the accuracy of the results: the more trees, the more accurate the result. However, it's important to highlight that making the decision with the information gain or gain index approach is not the same as producing the forest.

B. K-Nearest Neighbor

K-Nearest Neighbors is the simplest but critical class of algorithms in Machine Learning. It belongs to the supervised gaining knowledge of the area and unearths excessive utility in sample recognition, information mining and intrusion detection.

It is broadly disposable in real-existence eventualities because it's miles non-parametric, meaning, it does now no longer make any underlying assumptions approximately the distribution of information (in preference to different algorithms including GMM, which expect a Gaussian distribution of the given information).

C. Support Vector Machine

This is a supervised algorithm for gaining knowledge of the set of rules which may be used for types of regression challenges. However, it's far mainly utilized in type problems. In the SVM set of rules, we plot every information object as a factor in n-dimensional Space(wherein 'n' is the variety of capabilities) with the price of every characteristic being the price of a specific point. Then, we carry out type via way of means of locating the hyper-plane that separates the 2 instructions very well (examine the beneath image).

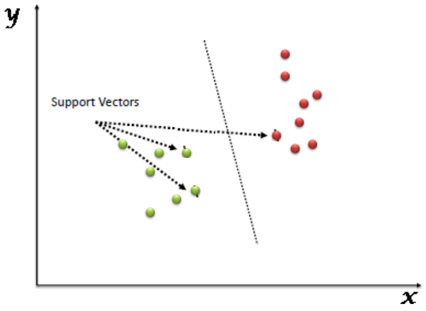


Fig 2. Support Vector Classifier scatter plot

Support Vectors are the points of a person's observation. The SVM classifier is a partition that separates the 2 instructions (hyper-plane/ line).

D. Decision Tree

The Decision Tree belongs to supervised algorithms. Unlike different supervised algorithms, it is used for fixing regression and category issues too.

The purpose of the usage of a Decision Tree is to create a version which can be used to expect the elegance or cost of the goal variable via means of studying easy choice regulations taken from previous data (educational data).

For predicting a category label for a document, we begin from the foundation of the tree. We examine the values of the foundation characteristic with the document's characteristics. On the premise of comparison, we observe the department like that cost and bounce to the following node.

E. Linear regression

Linear regression is more straightforward to apply and analyze, as well as to train. Overfitting is a common issue with linear regression; however, it may be readily prevented by making use of dimensionality reduction techniques and regularization techniques along with cross-validation.

All the five algorithms used above are part of an experiment to select the one with the best accuracy in predicting the severity of covid-19 from a patient's past health history data.

IV. LIMITATIONS

A. Random Forest

It's good at classification but not so much at regression because it can't forecast continuous nature well. Regression does not forecast beyond the range of the training data, and it is possible to overfit, especially for noisy data sets.

B. K Nearest Neighbour

This algorithm works well only with data not having large datasets and many dimensions. Feature scaling is required before applying the KNN algorithm or else it might generate wrong predictions. It is sensitive to noisy data, missing values, and outliers.

C. Linear Regression

This is liable to overfitting but can be prevented by using dimensionality reduction. The main limitation is that it assumes linearity for the dependent variable and the independent variable. It is also prone to multicollinearity because it thinks there is no relation among independent variables.

V. Testing

A. Exploratory data analysis

The graph below shows the different severity of the patients in the data set.

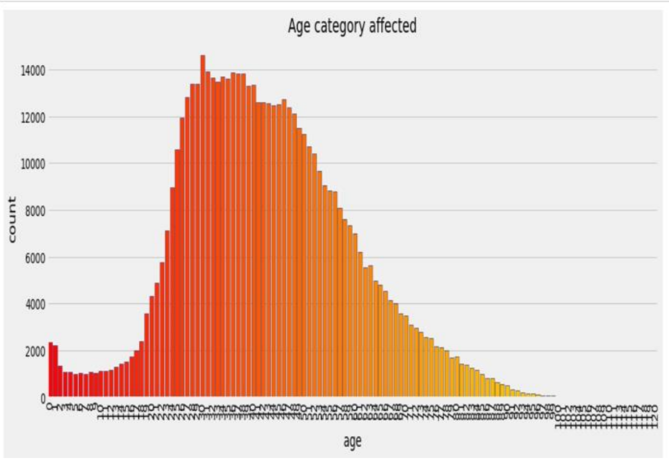
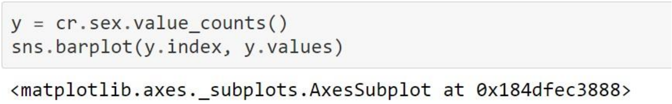


Fig 3.1 Severity of the patients w.r.t Age

The graph below shows the different sex of the patients in the data set.



```
y = cr.sex.value_counts()
sns.barplot(y.index, y.values)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x184dfec3888>
```

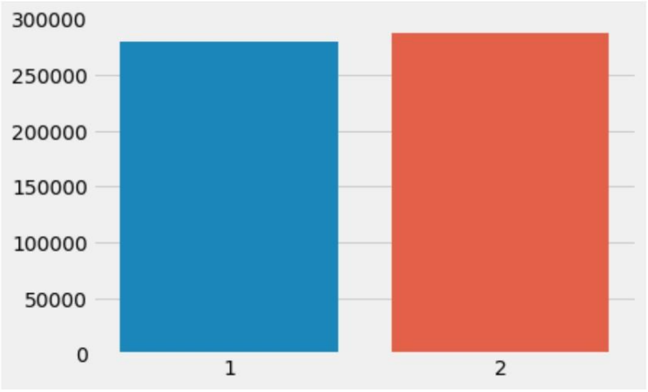
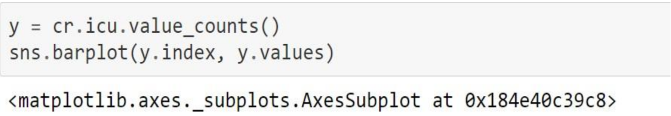


Fig 3.2 Bar plot on basis of the sex of the patients

The graph below shows the different severity of the patients in the data set.



```
y = cr.icu.value_counts()
sns.barplot(y.index, y.values)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x184e40c39c8>
```

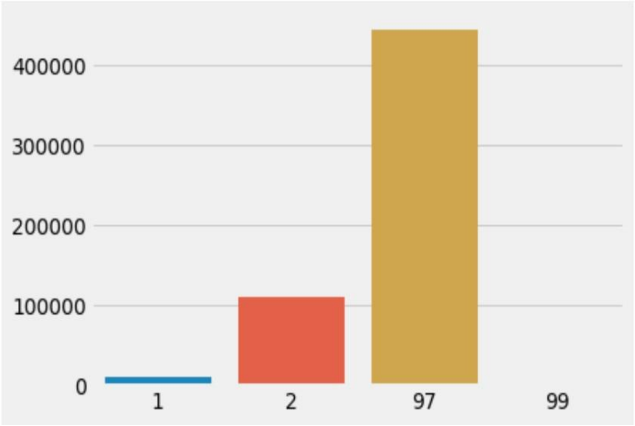


Fig 3.3 Different severity of patients

The graph below shows the feature importance of the classifier.

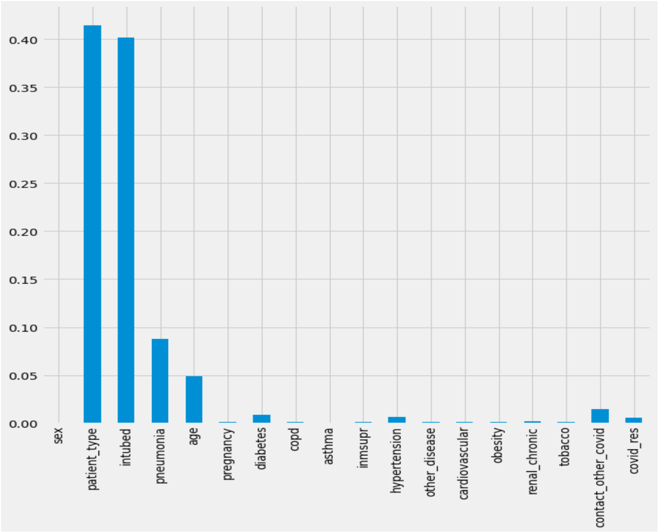


Fig 3.4 Feature importance from the classifier.

VI. RESULTS AND DISCUSSION

Algorithm	Accuracy
Random Forest	98.13
KNN	92.00
Linear Regression	98.82
Decision Tree Classifier	97.70
Support Vector Classifier	98.7

As it can be viewed in the above table, SVC scored the highest accuracy and KNN scored the least.

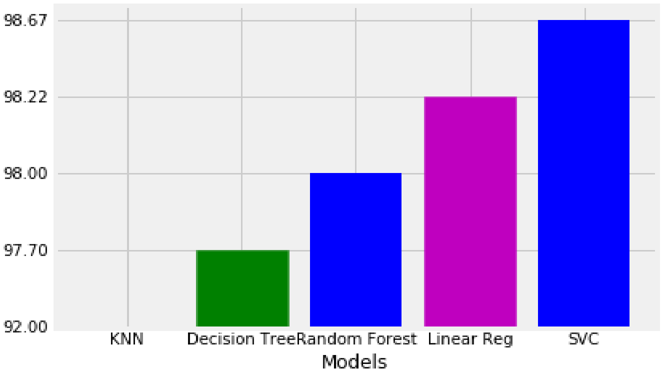


Fig 4.1 Results of test sets

VII. CONCLUSION

The proposed model is data-dependent and is subject to modification in the covid-19 virus. The model will be needing new data to understand patients accurately. The data set needs to be large with more features to play with so that we are sure it performs accurately. This model can be deployed with real-time data and used by doctors to classify the patients at the beginning. The model can be modified for other risk predictions of another disease.

VII. REFERENCES

[1] Sina F. Ardabili and Amir Mosavi, ‘’ COVID-19 Outbreak Prediction with Machine Learning “,ResearchGate publications,May 25 .2020.

[2] Furqan Rustam, Aijaz Ahmad Reshi, (Member, IEEE), Arif Mehmood 3, Saleem Ullah, Byung-Won, Waqar Psalm,” COVID-19 Future Forecasting Using Supervised Machine Learning Models “, IEEE Access, Vol. 8, May 5 .2020

[3] Zirun Zhao, Anne Chen, Wei Hou, James M. Graham, Haifang Li, Paul S. Richman, Henry C. Thode, Adam J. Singer, Tim Q. Duong “Prediction model and risk scores of ICU admission and mortality in COVID-19“, PLOSON, https://journals.plos.org/plosone/article/file?type=printable&id=10.1371/journal.pone.0236618 ,18 July 30, 2020.

[4] Roseline Oluwaseun OGUNDOKUN, Joseph Bamidele AWOTUNDE, “Machine Learning Prediction for COVID 19 Pandemic India”, https://www.medrxiv.org/content/10.1101/2020.05.20.20107847v1 ,May 26, 2020.

[5] Prathamesh Parchure, Himanshu Joshi, Kavita Dharmarajan, Robert Freeman, David L Reich, Madhu Mazumdar, rem Timsina, Arash Kia “Development and validation of a machine learning- based prediction model for near-term in-hospital mortality among patients with COVID-19 “, 18 August. 2020.

[6] Andre Patricio, Rafael S. Costa and Rui Henriques “COVID-19 in Portugal: predictability of hospitalization, ICU and respiratory assistance needs” , pp.1-8, https://www.researchgate.net/publication/345968499_COVID-19_in_Portugal_predictability_of_hospitalization_ICU_and_respiratory-assistance_needs

[7] Siddharth Singh, Piyush Raj, Raman Kumar and Rishu Chaujar “Prediction and forecast for COVID-19 Outbreak in India based on Enhanced Epidemiological Models “, IEEE Access, Second International Conference, pp .93-97, 978-1-7281-5374-2 ,2020.

[8] Jitian Li, Zhe Chen, Yifei Nie, Yan Ma, Qiaoyun Guo, Xiaofeng Dai, “Identification of Symptoms Prognostic of COVID-19 Severity: Multivariate Data Analysis of a Case Series in Henan Province”, JMIR Publications, Vol 22, No 6, June (2020). Available at: https://www.jmir.org/2020/6/e19636/

[9] The dataset used for the analysis is taken from www.kaggle.com.

[10] Ram Kumar Singh, Martin Drews, Manuel De La Sen, Manoj Kumar, Sati Shankar Singh, Ajai Kumar Pandey, Prashant Kumar Srivastava, Manmohan Dobriyal, Meenu Rani, Preeti Kumari, And Pavan Kumar, (Member, IEEE), “Short-Term Statistical Forecasts of COVID-19 Infections in India “, IEEE Access, vol 8, pp. 186932 – 186938, Oct 22.2020

[11] R. K. Singh, M. Rani, A. S. Bhagavathula, R. Sah, A. J. RodriguezMorales, H. Kalita, C. Nanda, S. Sharma, Y. D. Sharma, A. A. Rabaan, J. Rahmani, and P. Kumar, ‘‘Prediction of the COVID-19 pandemic for the top 15 affected countries: Advanced autoregressive integrated moving average (ARIMA) model,’’ JMIR Public Health Surveill., vol. 6, no. 2, May 2020, Art. no. e19115, doi: 10.2196/19115.

[12] Bayes C, Valdivieso L. “Modelling death rates due to COVID-19: a Bayesian approach”. arXiv. (2020) 2004.02386. Available online at: https://www.researchgate.net/publication/340475268_Modelling_death_rates_due_to_COVID-19_A_Bayesian_approach

[13] Sujatha R, Chatterjee JM, Hassanien AE. A machine learning forecasting model for COVID- 19 pandemic in India. Stoch Environ Res Risk Assess. (2020) 34:959–72. https://pubmed.ncbi.nlm.nih.gov/32837309/