Covid Detection using KNN Algorithm and Epidemic Prediction using ML Algorithm

Zaibunnisa L. H. Malik
Computer Engineering,
M. H. Saboo Siddik Polytechnic
Mumbai, India.

Shaikh Iqra
Computer Engineering,
M. H. Saboo Siddik Polytechnic
Mumbai, India.

Patel Shifa
Computer Engineering,
M. H. Saboo Siddik Polytechnic
Mumbai, India.

Shaikh Sania
Computer Engineering,
M. H. Saboo Siddik Polytechnic
Mumbai, India.

Abstract— **Generally, all factors associated with a disease outbreak poses an intractable challenge. Our plan is to research and verify the unfold of pandemic sicknesses in towns and sub-metropolitan areas, wherever care may not be without delay on the market. We would like to create an AI model that might predict the infectious disease dynamics. Our methodology takes into thought the geographics, climate and populace conveyance of an affected space, as these are relevant options and unobtrusively contribute to infectious disease dynamics. This research uses machine learning to predict these factors. The goal of creating this system is to aid in deciding whether an outbreak should get immediate attention or not. Almost all systems that predict COVID using clinical dataset having parameters and inputs from complex tests conducted in labs. None of the systems predicts COVID based on risk factors. Detecting COVID using various Risk Factors can result into a more accurate outcome and can help Medical Professionals to evaluate and treat their patients in a more productive manner.**

Keywords—COVID, AI, KNN, Flask

I. INTRODUCTION

Solid forecasts of sickness communicable disease dynamics will be valuable to general wellbeing associations that arrange interventions to diminish or stop disease transmission. With the massive information growth in care and medicine sector, correct analysis of such information may facilitate in early sickness detection and higher patient care. With the supply of giant procedure power at hand, it's currently substantially viable to use the 'huge information' for foreseeing and managing a plague, an endemic, a scourge, a pandemic, a lethal disease, a virus, a harmful disease, a pestilence flare-up. Our plan is to research and verify the unfold of pandemic sicknesses in towns and sub-metropolitan areas, wherever care may not be without delay on the market. we would like to create a AI model that might predict the infectious disease dynamics. Our methodology takes into thought the geographics, climate and populace conveyance of an affected space, as these are relevant options and unobtrusively contribute to infectious disease dynamics. Our model would be useful for the care authorities by aiding them take the acceptable action in wording of reassuring that enough resources are on the market to satisfy the necessity and, if potential, curb the prevalence of such infectious disease. curb the preventable disease-relatedaffliction.

Minimize monetary burden on governments and medical services frameworks by giving them first-hand data concerning irruption prone territories and tributary agents for the unfold of plague. Given a district wherever a plague, an endemic, a scourge, a pandemic, a destructive disease, a virus, a harmful disease, a pestilence flare-up has happened, our model ought to be ready to establish next irruption prone territories and establish options that contribute considerably within the unfold of the pandemic. Scourges of communicable disease are typically caused by many factors as well as a modification within the ecology of the host population. modification within the infectious agent reservoir or the presentation of AN rising infectious agent to a bunch population.

Physically representing all factors related to a sickness irruption poses AN disobedient challenge. This analysis uses AI to anticipate these components. The goal of making this method is to help when deciding whether or not a virus ought to get quick consideration or not. the majority systems that foresee COVID victimization clinical dataset having boundaries and sources of info fromadvanced tests led in labs. None of the frameworks predicts COVID supported risk factors. police investigation COVID victimization varied Risk Factors may end up into an additional correct outcome and might facilitate Medical Professionals to gauge and treat theirpatients in a very additional productive way.

II. LITERATURE SURVEY

Reliable predictions of infectious disease dynamics can be valuable to public health organizations that plan interventions to decrease or prevent disease transmission. With the big data growth in healthcare and biomedical sector, accurate analysis of such data could help in early disease detection and better patient care. With the availability of huge computational power at hand, it is now very much viable to exploit the 'big data' for predicting and managing an epidemic outbreak. Our

Special Issue - 2021

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
NREST - 2021 Conference Proceedings

idea is to analyse and determine the spread of epidemic diseases in villages and sub-urban areas, where healthcare might not be readily available. We want to build a machine learning model that could predict the epidemic disease dynamics. Our approach takes into consideration the geography, climate and population distribution of an affected area, as these are relevant features and subtly contribute to epidemic disease dynamics. Our model would be beneficial for the healthcare authorities by assisting them take the appropriate action in terms of assuring that enough resources are available to suffice the need and, if possible, curbing the occurrence of such epidemic disease. Curbing the preventable disease-related suffering. Minimize financial burden on governments and health care systems by providing them first-hand information about outbreak prone areas and causative agents for the spread of epidemic. Given an area where an epidemic outbreak has occurred, our ML model should be able to identify next outbreak prone areas and identify features which contribute significantly in the spread of the outbreak. Epidemics of infectious disease are generally caused by several factors including a change in the ecology of the host population. Change in the pathogen reservoir or the introduction of an emerging pathogen to a host population. The feature vectors in our model are general enough to be adapted with a slight change to study any epidemic disease. Detecting COVID using various Risk Factors can result into a more accurate outcome and can help Medical Professionals to evaluate and treat their patients in a more productive manner. Manually accounting for all factors associated with a disease outbreak poses an intractable challenge. This research uses machine learning to predict these factors. The goal of creating this system is to aid in deciding whether an outbreak should get immediate attention or not. Almost all systems that predict COVID using clinical dataset having parameters and inputs from complex tests conducted in labs. None of the systems predicts COVID based on risk factors. Detecting COVID using various Risk Factors can result into a more accurate outcome and can help Medical Professionals to evaluate and treat their patients in a more productive manner..

III. PROPOSED SYSTEM

The Proposed System is developed using Python, with the help of web framework Flask. The system is made capable of detecting, evaluating and measuring the severity of an outbreak using some predefined parameters which are taken in count. The system comprises of two modules: Epidemic and COVID module. The data collected by these modules are sent to the ML Algorithm which thereafter evaluate the inputs and predicts the result. The following diagram shows the complete System Architecture:
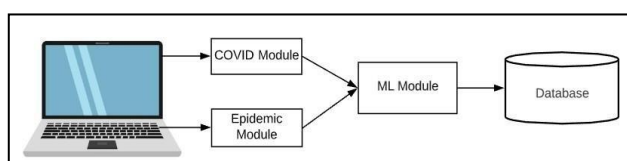

Fig. System Architecture Diagram

A. Support Vector Classifier (SVC) Algorithm

Support Vector Classifier Algorithm's are powerful yet flexible supervised machine learning algorithms which are used both for classification and regression. But generally, they are used in classification problems. The Epidemic Module collects Data input from the user and sent the data to Support Vector Classifier (SVC) Algorithm. The SVC Algorithm is assigned the work to evaluate the data and generate the result. The accuracy of Support Vector Classifier Algorithm is 75%

B. K-Nearest Neighbour (KNN) Algorithm

K Nearest Neighbour (KNN) Algorithm is a very simple, easy to understand, versatile and one of the topmost machine learning algorithms. KNN used in the variety of applications. We have KNN Algorithm in COVID Prediction module. The data provided to the algorithm in turn is used to find the nearest identical data in the dataset. Based on the evaluation, the result is generated.

C. Dataset

The dataset used in KNN Module is referred from figshare posted by Yanyan Xu. The dataset provides factors that are used to evaluate COVID. The dataset has 90 rows with 16 parameters. The TABLE I below shows the identified important risk factors and the corresponding values and their encoded values in brackets, which were used as input to the system.

	Factors	Values
1	Gender	Male (1), Female (0)
2	Age (years)	20-34 (-2), 35-50 (-1), 51-60 (0), 61-79 (1), >79 (2)
3	No comorbidity	Yes (1) or No (0)
4	Cardiovascular & cerebrovascular disease	Yes (1) or No (0)
5	Endocrine system disease	Yes (1) or No (0)
6	Malignant tumour	Yes (1) or No (0)
7	Respiratory system disease	Yes (1) or No (0)
8	Digestive system disease	Yes (1) or No (0)
9	Renal disease	Yes (1) or No (0)
10	Liver disease	Yes (1) or No (0)
11	Fever	Yes (1) or No (0)
12	Cough	Yes (1) or No (0)
13	Chest tightness	Yes (1) or No (0)
14	Fatigue	Yes (1) or No (0)
15	Diarrhoea	Yes (1) or No (0)
16	RNA clearance	Yes (1) or No (0)
Output	COVID	Yes (1) or No (0)

Table.1 Risk Factors Values and Encodings.

Data analysis has been carried out in order to transform data into useful form, for this the values were encoded mostly between a range [-1, 1]. Data analysis also removed the inconsistency and anomalies in the data. This was needed. Data analysis was needed for correct data pre-processing. The removal of missing and incorrect inputs will help the algorithm to generalize well.

Special Issue - 2021

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
NREST - 2021 Conference Proceedings

Almost all systems that predict COVID using clinical dataset having parameters and inputs from complex tests conducted in labs. None of the systems predicts COVID based on risk factors. Our model would be beneficial for the healthcare authorities by assisting them take the appropriate action in terms of assuring that enough resources are available to suffice the need and, if possible, curbing the occurrence of such epidemic disease.

- Minimize financial burden on governments and health care systems by providing them first-hand information about outbreak prone areas and causative agents for the spread of epidemic.
- Given an area where an epidemic outbreak has occurred, our ML model should be able to identify next outbreak prone areas and identify features which contribute significantly in the spread of the outbreak.
- Epidemics of infectious disease are generally caused by several factors including a change in the ecology of the host population.
- Change in the pathogen reservoir or the introduction of an emerging pathogen to a host population.
- The feature vectors in our model are general enough to be adapted with a slight change to study any epidemic disease.
- Detecting COVID using various Risk Factors can result into a more accurate outcome and can help Medical Professionals to evaluate and treat their patients in a more productive manner.

IV. CONCLUSION AND FUTURE ASPECTS

The proposed system uses Risk Factors as input and evaluates and predicts the result based on these factors. If identified risk factors are less accurate then it may lead to false result. The accuracy of the result depends on the accuracy of the algorithms used. If the accuracy at any time decreases, the result generated by the algorithm also becomes less accurate. In Future, a survey can be done to increase the integrity and correctness of the dataset. By achieving, this we can train our model, which in turn can increase the accuracy of our algorithms.

The key limitation of the study is that we don't have a large set of data to train the prediction model. In future, our goal is to collect more data from different sources and to improve the prediction model accuracy. Another limitation of the application is that it is not suitable for the age group below 20 years of age as they are not included in the dataset.

REFERENCES

[1] Remuzzi, A.; Remuzzi, G. COVID-19 and Italy: what next? Lancet 2020.

[2] Ivanov, D. Predicting the impacts of epidemic outbreaks on global supply chains: A Transp. Res. Part E Logist. Transp. Rev. 2020, 136, doi:10.1016/j.tre.2020.101922.

[3] Koolhof, I.S.; Gibney, K.B.; Bettiol, S.; Charleston, M.; Wiethoelter, A.; Arnold, A.L.; Campbell, P.T.; Neville, P.J.; Aung, P.; Shiga, T., et al. The forecasting of dynamical Ross River virus outbreaks: Victoria, Australia. Epidemics 2020, 30, doi:10.1016/j.epidem.2019.100377.

[4] Darwish, A.; Rahhal, Y.; Jafar, A. A comparative study on predicting influenza outbreaks using different feature spaces: application of influenza-like illness data from Early Warning Alert and Response System in Syria. BMC Res. Notes 2020, 13, 33, doi:10.1186/s13104020-4889-5.

[5] Rypdal, M.; Sugihara, G. Inter-outbreak stability reflects the size of the susceptible pool and forecasts magnitudes of seasonal epidemics. Nat. Commun. 2019, 10, doi:10.1038/s41467-019-10099-y.

[6] Scarpino, S.V.; Petri, G. On the predictability of infectious disease outbreaks. Nat. Commun. 2019, 10, doi:10.1038/s41467-019-08616-0.

[7] Zhan, Z.; Dong, W.; Lu, Y.; Yang, P.; Wang, Q.; Jia, P. Real-Time Forecasting of Hand-Foot-and-Mouth Disease Outbreaks using the Integrating Compartment Model and Assimilation Filtering. Sci. Rep. 2019, 9, doi:10.1038/s41598-019-38930-y.

[8] Koike, F.; Morimoto, N. Supervised forecasting of the range expansion of novel non-indigenous organisms: Alien pest organisms and the 2009 H1N1 flu pandemic. Global Ecol. Biogeogr. 2018, 27, 991-1000, doi:10.1111/geb.12754.

[9] Dallas, T.A.; Carlson, C.J.; Poisot, T. Testing predictability of disease outbreaks with a simple model of pathogen biogeography. R. Soc. Open Sci. 2019, doi:10.1098/rsos.190883.

[10] de Groot, M.; Ogris, N. Short-term forecasting of bark beetle outbreaks on two economically important conifer tree species. For. Ecol.Manage.2019, 450, doi:10.1016/j.foreco.2019.117495.