Dynamic Analytics and Forecasting Model for Covid-19 Using Machine Learning Algorithms

C. Siva
Associate Professor, Department of Information Technology, Nandha Engineering College, Erode, India.
E-mail: sivachelladurai@gmail.com

K.G. Maheshwari
Assistant Professor (Sr), Department of Information Technology, Government Engineering College, Erode, India.
E-mail: kgmaheswari@gmail.com

G. Nalinipriya
Professor, Department of Information Technology, Saveetha Engineering College, Chennai, India.
E-mail: nalini.anbu@gmail.com

J. Priscilla Mary
Assistant Professor, Department of Information Technology, Nandha College of Pharmacy, Erode, India.

Received June 10, 2021; Accepted September 12, 2021
ISSN: 1735-188X
DOI: 10.14704/WEB/V18SI05/WEB18302

Abstract

In our day to day life, the availability of correctly labelled data as well as handling of categorical data are mostly acknowledged as two main challenges in dynamic analysis. Therefore, clustering techniques are applied on unlabelled data to group them in accordance with the homogeneity. There are many prediction methods that are being popularly used in handling forecasting problems in real time environment. The outbreak of coronavirus disease (COVID19)-2019 creates the need for a medical emergency of worldwide concern with a rapidly high danger of open out and strike the entire world. Recently, the ML prediction models were used in many real time applications which necessitate the identification and categorization for real time environment. In medical field Prediction models are vital role to obtain observations of spread and significances of infectious diseases. Machine learning related forecasting mechanisms have showed their importance to develop the decision making on the upcoming course of actions. The K-means algorithm and hierarchy were applied directly on the renewed dataset using R programming language to create the covid patient cluster. Confirmed Covid patients count are passed to Prophet package, then the prophet model has been created. This forecasts model predicts the future covid count, which is

essential for the clinical and healthcare leaders to make the appropriate measures in advance. The results of the experiments indicate that the quality of Hierarchical clustering outperforms than the K-Means clustering algorithm in the structured structured dataset. Thus, the prediction model also used to support model predictions help for the officials to take timely actions and make decisions to contain the COVID-19 dilemma. This work concludes Hierarchical clustering algorithm is the best model for clustering the covid data set obtained from world health organization (WHO).

Keywords

COVID 19, Forecasting Model, Machine Learning Algorithms, Dynamic Analytic, k Means Clustering, Hierarchical Clustering.

Introduction

Covid-19 is a very thoughtful deadly disease that has been announced as a widespread by the world health organization (WHO) (Velavan, T.P., 2020). The entire world is waiting with all its might to end Covid-19 pandemic, which will affect the countries in many ways like health education and economic problems (Amato, A., 2020). During the current common insistence, experts, clinicians, and healthcare experts around the world keep on penetrating for a new expertise to support in tackling the Covid-19 pandemic (Keesara, S., 2020). The performance of Machine Learning (ML) and Artificial Intelligence (AI) application on the earlier epidemic inspire researchers by giving a novel approach to fight against the novel Coronavirus epidemic (Zhang, P., 2021). COVID-19 is an emerging disease that can give rise to respiratory problems and pneumonia (Lepri, G., 2020). Clinical symptoms that seem vary, ranging from symptoms such as the common cold, cough, running nose, pharyngitis, muscle aches, headache to severe complications (Eccles, R., 2014). Predicting the out spread of coronavirus and finding distribution of COVID cases across the countries is the most challenging thing in recent days (Tuli, S., 2020). Estimating the pandemic with high precision will help major countries make ready with a plan to confront a war against the virus spread (Malki, Z., 2021).

Related Works

(Doroshenko, A, 2020) Form clusters for provinces in Italy K- Means and Hierarchical and it was spotted that the clusters acquired collaborate regions with a related level of industrial growth. Therefore, the finest epidemiological circumstance is noticed in regions with badly evolved economies or in those where the major areas of occupation are:

agricultural, construction along with commerce. The worst is in areas with large industry as well as other industries.

(Gupta, V.K., 2021) identified the COVID-19 cases - positive, mortal, and healed cases in India alone and carried out this analysis based on the cases come about in various states of India in sequential dates. The dataset holds different classes so multi-class grouping is performed. In this dataset, data cleansing and attribute selection have applied and then carried out predicting of all classes with the use of decision tree, linear model, support vector machine (SVM), random forest, together with neural network, where random forest model surpassed the others, and hence, concluded that the random forest is used for forecasting and survey of all the results.

(Bhati, A., 2020) reported the Bhilwara model and differentiate the model with India COVID-19 pandemic lockdown along with a forecast for a decrease in the number of future cases with its execution. The Bhilwara model is imitated with the help of 3^{rd} - degree polynomial curve fitting techniques, and the mean successive rate of infection on the basis of the COVID-19 spread curve for a certain group of days which illustrate the outcome of policies established by Bhilwara administration. Using calculated mean successive rate, COVID-19 spread is forecast with 3rd -degree polynomial regression using a dataset of each and every states of India.

(Kurniawan, R., 2020) chosen K-Means clustering and interconnection on a Coronavirus Outbreak COVID-19 data collected on March 27 and August 16, 2020. The K-means naturally search for unknown clusters of several countries infected with the COVID19 enormously. It shows that a group of m=5 generate a precision of about 97% with the [United States and Italy], [Iran, France], [Spain, German], [Indonesia, Malaysia, Philippine] as groups. Simultaneously, it predicts a relevant relationship between the total mortals and serious patients' attributes of 0.85 while interconnecting COVID-19 characteristics.

(Ibrahim, A.M., 2020) Depends upon the population formed the groups and each group virus out spread modelled in MATLAB. The modelling is carried out on the basis of exponential growth equation. The stringency index model was also used as an origin of study with respect to the government reaction of the groups in analysis. The population density was found to be not an important donor in controlling Covid-19 pandemic in the very foremost month of its widespread (Kandasamy, R., 2015).

(Darapaneni, N., 2020) chosen Autoregressive integrated moving average, for time series data forecast with 95% conviction. Modified logistic development model, where the study was done on data prior to the lockdown was raised, we found a finer AIC and R-squared values. The usual cause for the observation could be due to a slower development rate of the infection in the country at the time of lockdown showing a considerable advancement in the prediction.

Methodology and Data Analysis

This forecasting system used to improve the prediction might be very much useful in decision making to control the current scenario to guide interferes to control the issue in an effective manner. Machine Learning algorithms are used here to build systematic models. These systematic models are used in forecasting to obtain perceptions of spread and significances of infectious diseases. On the basis of previous research work, two different Machine Learning algorithms were implemented on dataset. R programming tool is used to detect the chances of covid-19 analysis. System architecture methods of the proposed system as well as the workflow as mentioned below:

1. Proposed System Architecture

The fig 1 shows the architectural diagram of proposed system. The first step is Data Importing and Data Pre-processing. In data importing, data have to be loaded in to the R environment for analysis.

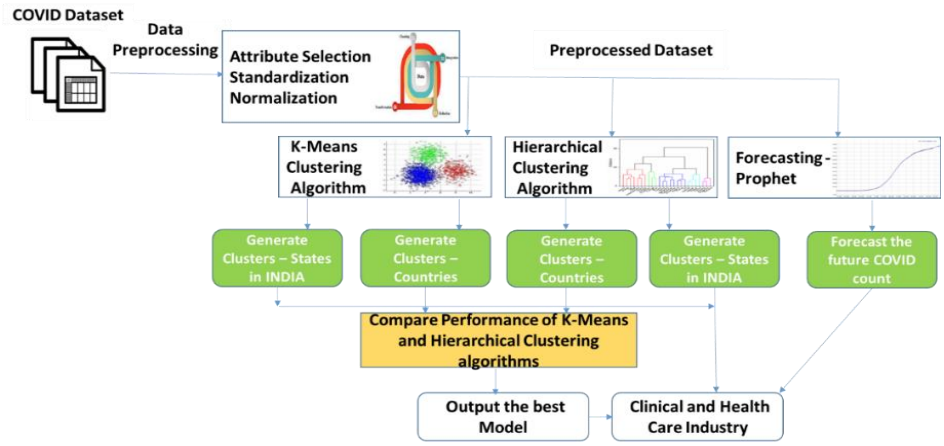

Fig. 1 The architectural diagram of proposed system

In data pre-processing, attribute selection, standardization and normalization functions will be applied. In standardization, raw data is transformed into common, understandable format. In attribute selection, hold only the attributes which is affecting the analysis and it

is not necessary to hold all the attributes for doing the analysis. In Normalization, mean of the attribute will be 0 and standard deviation will be 1. The pre-processed data is passed to the K-Means clustering algorithm. K-Means algorithm form the clusters among the countries in the world using the COVID patient's count. Hierarchical clustering algorithm formed the clusters among the countries. The performance of both algorithms is compared in terms of quality of the formation of cluster. For the quality measurements, components such as between sum of squares, within sum of squares, total sum of squares and agglomerative coefficient have taken into account. Using Prophet, the future covid count for India is calculated. The hierarchical clustering Model have formed a quality cluster when compared to K-means for this world health organization's COVID data set. So, this model along with the COVID patient's count will be given to the clinical and health care industry to take proper precautions in advance.

2. Machine Learning Approach

Machine learning principals are smart approaches used to predict, analysis and improve the performance of the system using supervised and unsupervised learning method. Recently machine learning based predication models are developed for real time applications like health care industry, share market, business analysis, education etc. based on the preceding research work, K-means algorithm and hierarchical algorithm-based clustering Model are implemented in COVID dataset. Then, how machine learning techniques are integrated with prediction future count of COVID patients with all states in India.

- ### K-Means Algorithm

K means algorithm is an iterative algorithm that attempts to group the given data into K pre-defined different non-overlapping subclasses (clusters) and each data point fit in to only one collection. The group of similar objects form same groups and unrelated objects belong to another cluster (Vijayakumar, J., 2013).

Algorithmic steps for k-means clustering
input
X ← ***set of data point x= {x_1,x_2…x_n}***
V ← ***Set of cluster center V= {v_1, v_2,…v_n}***
Repeat
calculate ← ***Distance of datapoint and cluster center***
Do
For i to N do
Minimum ← ***Assign datapoint of cluster center***
Compute ← ***New cluster center***
$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$
While(i<0)
Output
c ← ***cluster center***

Using K-Means algorithm, clustering model has been created by taking world health organization datasets as input for clustering countries. This Clustering models are generating the output of 3 sets of clusters, first cluster have been formed with the set of states with higher covid count, second cluster have been formed with the set of states with lesser covid count and third cluster have been formed with the set of countries with medium number of covid counts.

- **Hierarchical Algorithm**

A Hierarchical clustering method based on tree structure of grouping data as clusters. In this method each data points as a distinct cluster. This model produces a hierarchical sequence of nested clusters. A Dendrogram diagram used to represent, the sequence of order in which factors are merged and splits.

Algorithmic steps for Hierarchical clustering algorithm:
Input X ← ***each point as cluster point***
Calculate ← ***distance between two input cluster point***
Cluster ← ***each point as cluster***
Repeat
Cluster ← ***set*** the closest pair of clusters and merge into a single cluster
update ← ***update the distance between the cluster***
Until ← ***single cluster***

Using hierarchical clustering algorithm, clustering model has been created by taking world health organization datasets as input. Clustering model is generating the output of 3 sets of clusters, first cluster have been formed with the set of states with higher covid count, second cluster have been formed with the set of states with lesser covid count and third cluster have been formed with the set of countries with medium number of covid counts.

3. Work Flow of Proposed System

The figure 2 shows the work flow of the proposed system. In this work, dataset covid -19 was implemented and after apply the pre-processing, the renewed dataset was implemented in this work. Here based on the machine learning algorithm K means and Hierarchical algorithm, dataset give the three types of algorithm output gives the Higher, medium and lower covid count country.

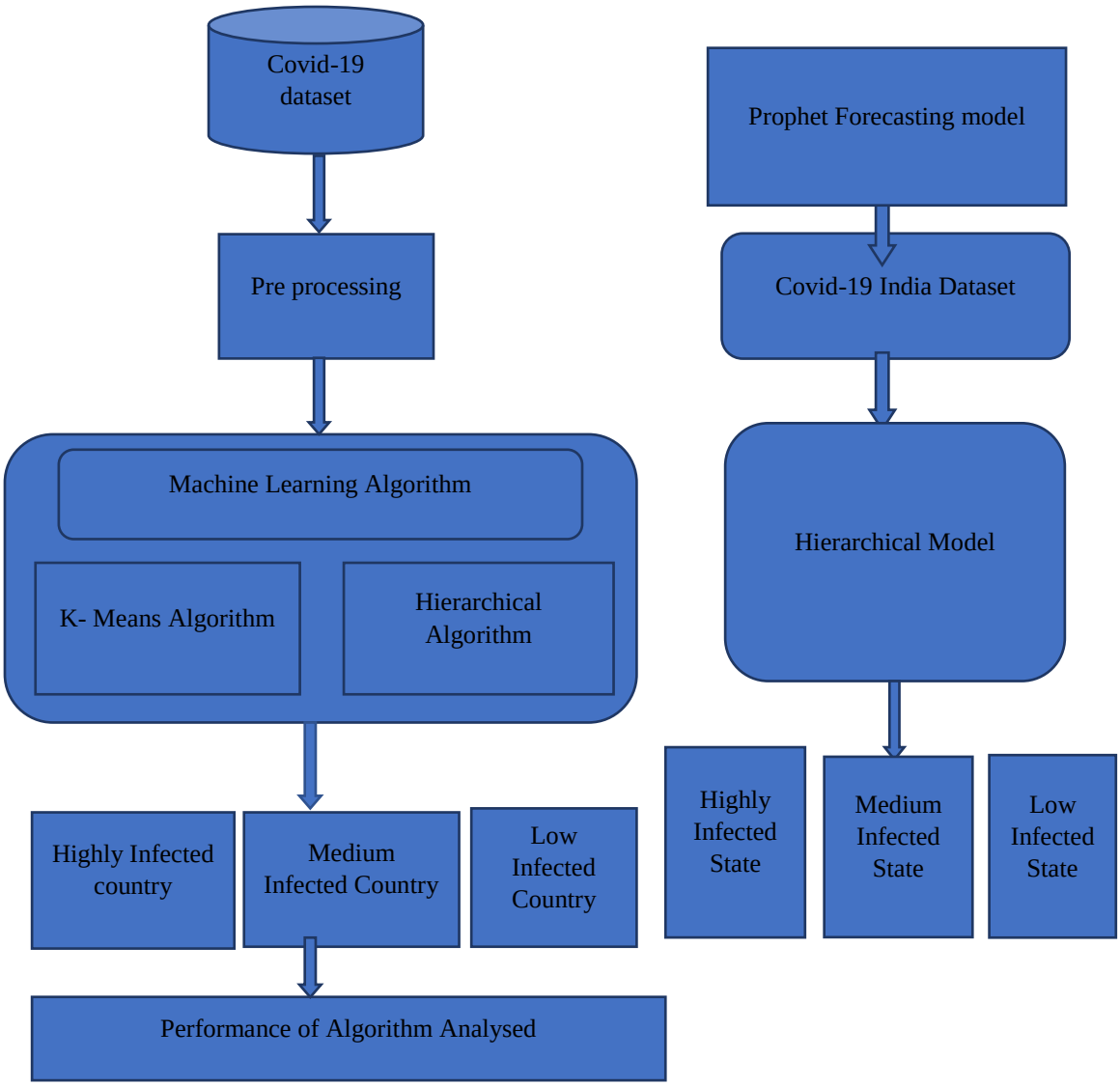

Fig. 2 Work flow of proposed system

Based on the performance of analysed the best performance algorithm was implemented on the India coid-19 dataset. The output of the algorithm gives the highly infected state, medium infected state and low infected state. prophet forecasting model was implemented the all the states in the India.

Use case diagram: The Use Case Diagram contains relationships between actors and instances. It signifies all the scenarios, related with proposed systems. Fig 3 shows the use case diagram of proposed system.

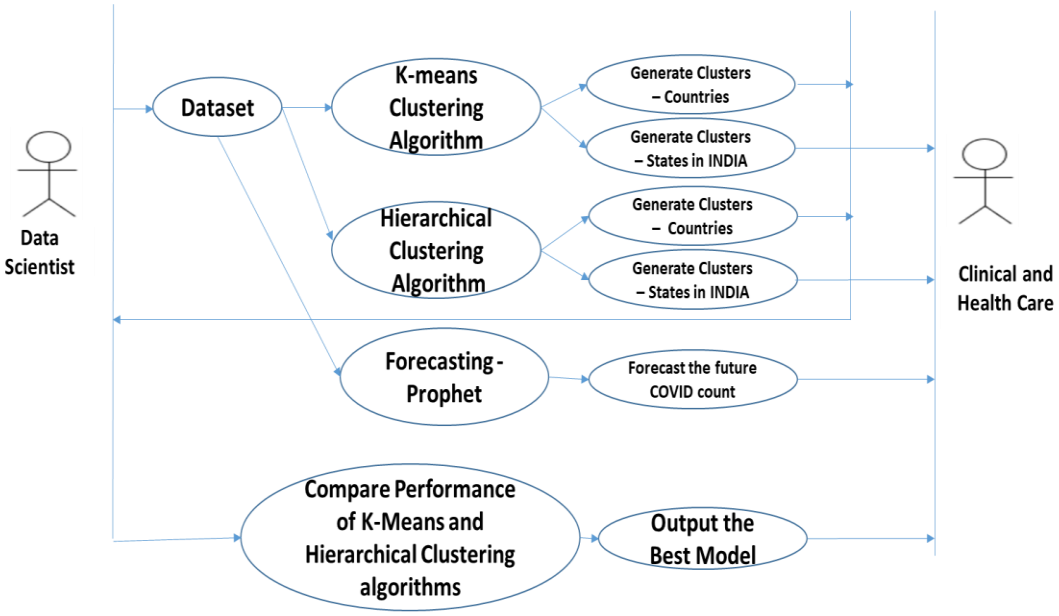

Fig. 3 Use case diagram of proposed system

Sequence Diagram: Sequence Diagrams are also so-called interaction diagrams, which will give the activity of the system. Fig 4 illustrates the sequence diagram of proposed system.

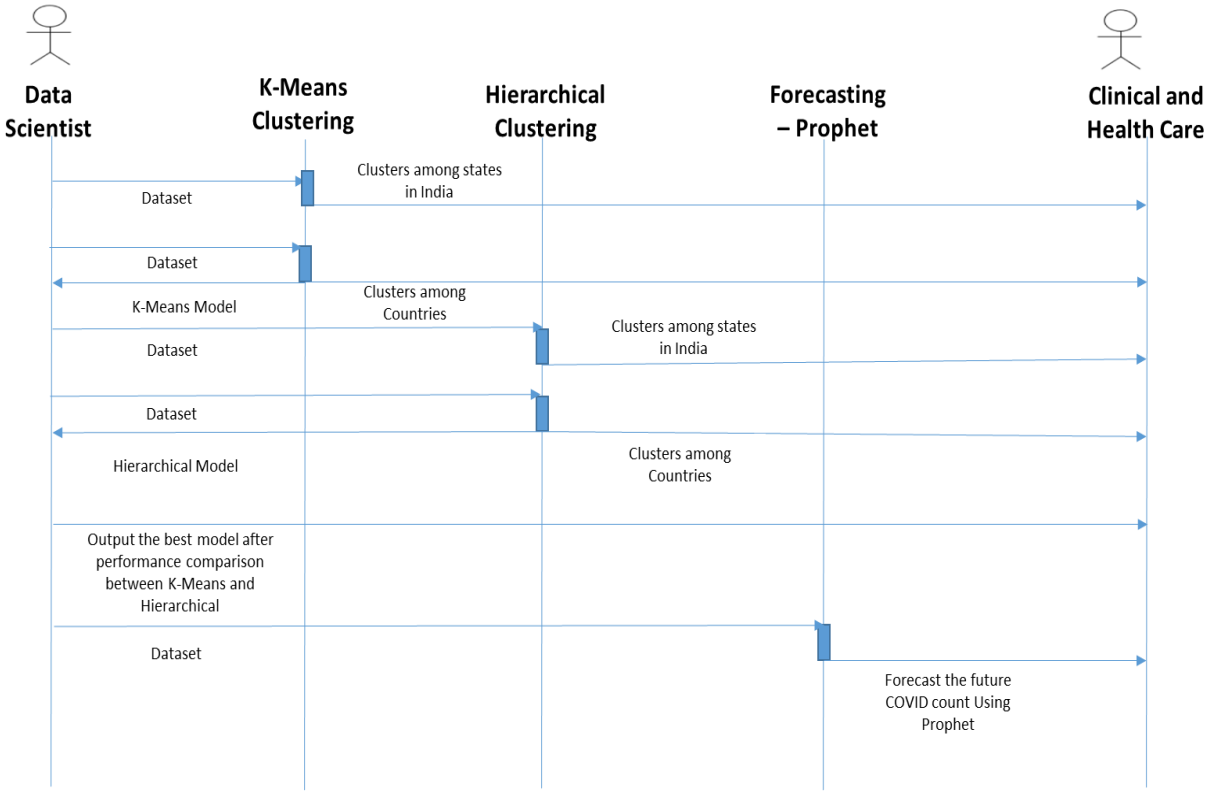

Fig. 4 Sequence diagram of the proposed system

Results and Discussion

In this section, the performance of the proposed hierarchical classifier is evaluated and compared. Covid 19 dataset have been utilized for the evaluation purpose. The evaluations are made in R tool with the performance parameters namely accuracy, precision, recall, and f-measure.

1. Dataset Description

2019 Novel Coronavirus (2019-nCoV) is a virus (more specifically, a coronavirus) identified as the cause of an outbreak of respiratory illness first detected in Wuhan, China. Early on, many of the patients in the outbreak in Wuhan, China reportedly had some link to a large seafood and animal market, suggesting animal-to-person spread. However, a growing number of patients reportedly have not had exposure to animal markets, indicating person-to-person spread is occurring. At this time, it's unclear how easily or sustainably this virus is spreading between people – CDC. This dataset has daily level information on the number of affected cases, deaths and recovery from 2019 novel coronavirus. Please note that this is a time series data and so the number of cases on any given day is the cumulative number. The data is available from 22 Jan, 2020 (kaggle.com).

Main file in this dataset is covid_19_data.csv and the detailed descriptions are below.

- S no - Serial number.
- Observation Date - Date of the observation in MM/DD/YYYY.
- Province/State - Province or state of the observation (Could be empty when missing).
- Country/Region - Country of observation.
- Last Update - Time in UTC at which the row is updated for the given province or country. (Not standardised and so please clean before using it).
- Confirmed - Cumulative number of confirmed cases till that date.
- Deaths - Cumulative number of of deaths till that date.
- Recovered - Cumulative number of recovered cases till that date.

2. Numerical Analysis

Here comparison analysis forecasting model has been made for COVID 19 dataset gathered from different countries such as India, South korea, Brazil, and USA. The

numerical analysis is made in terms of accuracy, precision, recall and f-measure. The numerical values obtained are shown in the following table 1.

Table 1 Comparison of classifier algorithms

Performance metrics	**Countries**							
	India		**South korea**		**Brazil**		**USA**	
	K means	**Hierarchical**	**K means**	**Hierarchical**	**K means**	**Hierarchical**	**K means**	**Hierarchical**
Precision	71.43	93.3	71	89	82	87	79	85
Recall	100	100	98	98.1	87	91	83	87
F-measure	85.7	96.6	87	92	75	89	87	91
Accuracy	92	96	93	95	89	93	92	95

Precision: Precision is demonstrated as the ratio of the true positives opposite to both true positives and false positives result for imposition and real features. It is different as provided below.

$$\text{Precision(P)} = \frac{T_p}{T_p + F_p} \quad (1)$$

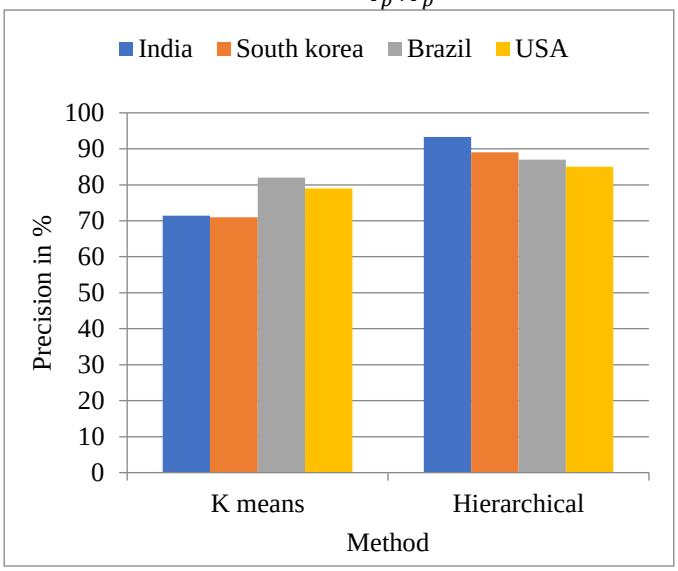

Figure 5 Precision comparison

From the above Figure 5, it can be noticed that the proposed hierarchical clustering system has higher precision when distinguished with other methods. The numerical analysis proved that the proposed Hierarchical shows 21.87% increased precision rate for India dataset, 18% increased precision rate for south korea dataset, 5% increased precision rate for brazil dataset and 6% increased precision rate for USA dataset.

Recall: Recall value is computed on the root of the data retrieval at true positive forecast, false negative. Generally, it can be determined as follows,

$$\text{Recall(R)} = \frac{T_p}{T_p + F_n} \quad (2)$$

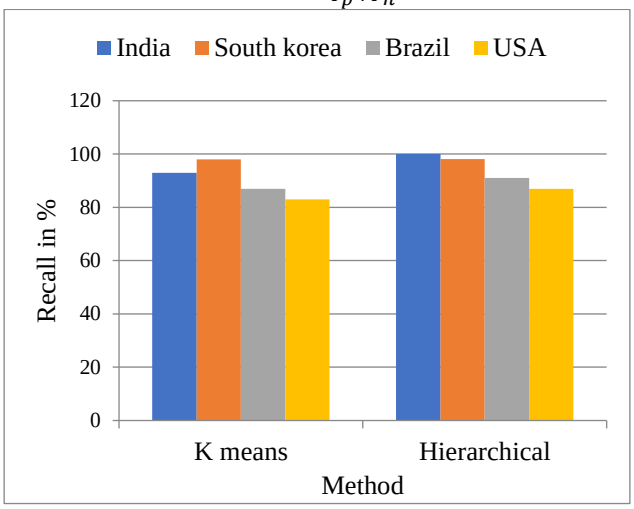

Figure 6 Recall comparison

From the above Figure 6, it can be noticed that the proposed hierarchical clustering system has higher recall when distinguished with other methods. The numerical analysis proved that the proposed Hierarchical shows 7% increased recall rate for India dataset, 0.1% increased recall rate for south korea dataset, 4% increased recall rate for brazil dataset and 4% increased recall rate for USA dataset.

F-measure: It is a measure of an accuracy of the test. It assume both the precision p and the recall *r* of the test to measure the score.

$$\text{F-measure} = 2.\frac{P.R}{P+R} \quad (3)$$

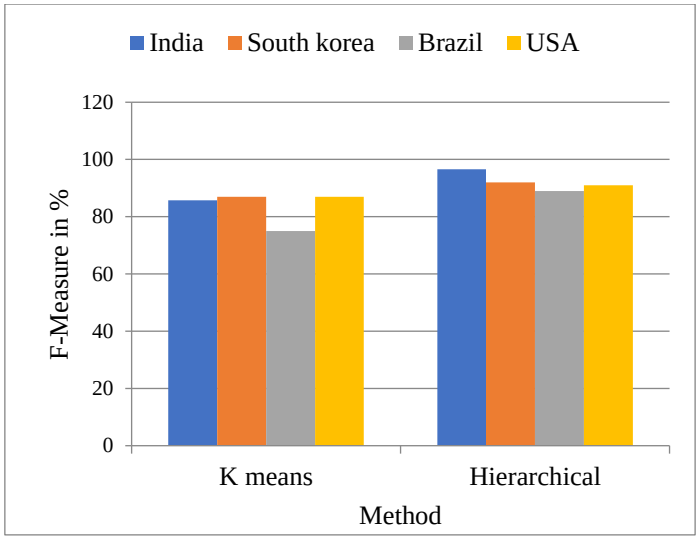

Figure 7 F-measure comparison

From the above Figure 7, it can be noticed that the proposed hierarchical clustering system has higher f-measure when distinguished with other methods. The numerical analysis proved that the proposed Hierarchical shows 10.9% increased f-measure rate for India dataset, 5% increased f-measure rate for south korea dataset, 14% increased f-measure rate for brazil dataset and 4% increased f-measure rate for USA dataset.

Accuracy: Accuracy is determined as the complete correctness of the model and is computed as the total actual classification parameters ($T_p + T_n$) which is classified by the sum of the classification parameters ($T_p + T_n + F_p + F_n$). The accuracy is calculated as like:

$$\text{Accuracy} = \frac{T_p + T_n}{(T_p + T_n + F_p + F_n)} \quad (4)$$

Where T_P is known as the amount of correct predictions that an instance is negative, T_n is termed as the amount of incorrect predictions that an instance is positive, F_p is known as the amount of incorrect of predictions that an instance negative, and F_n is known the amount of correct predictions that an instance is positive.

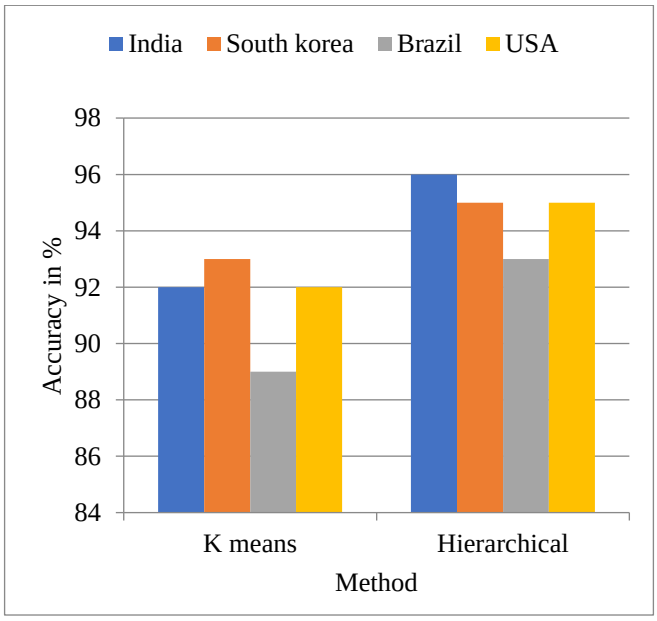

Figure 8 Accuracy comparison

From the above Figure 8, it can be noticed that the proposed hierarchical clustering system has higher accuracy when distinguished with other methods. The numerical analysis proved that the proposed Hierarchical shows 4% increased accuracy rate for India

dataset, 2% increased accuracy rate for south korea dataset, 4% increased accuracy rate for brazil dataset and 3% increased accuracy rate for USA dataset.

Conclusion

The core strategy of the project to find the groups of countries which is having higher, lower and medium COVID count among the countries in the world and among the states in India. Using hierarchical clustering and K-Means clustering algorithm, clusters have been created. The results of the experiments indicate that the quality of Hierarchical clustering outperforms the K-Means clustering algorithm. It also presented the trends and forecasts of future covid count in India. The Hierarchical clustering algorithm is the best model for clustering the covid data set obtained from world health organization. This Hierarchical clustering model and along with the future covid count is very essential for the clinical and healthcare leaders to make the appropriate measures in advance. Thus, the burden of the health sector can be reduced by foreseeing future situations and making strategies and plans. The future enhancement will focus on, this work takes only structured dataset as input, this can be improved by using unstructured dataset. Data can be tested with some more clustering algorithms, so that strong clusters can be obtained. In this work, time series forecasting prophet approach is applied, in future some more time series algorithm can be included to get optimum result. the algorithm, so that performance of the algorithm will be increased.

References

Velavan, T.P., & Meyer, C.G. (2020). The COVID-19 epidemic. *Tropical medicine & international health*, *25*(3), 278-280.

Amato, A., Caggiano, M., Amato, M., Moccia, G., Capunzo, M., & De Caro, F. (2020). Infection control in dental practice during the COVID-19 pandemic. *International journal of environmental research and public health*, *17*(13), 4769.

Keesara, S., Jonas, A., & Schulman, K. (2020). Covid-19 and health care's digital revolution. *New England Journal of Medicine*, *382*(23), e82.

Zhang, P., Wang, C., Kumar, N., Jiang, C., Lu, Q., Choo, K.K.R., & Rodrigues, J.J. (2021). Artificial Intelligence Technologies for COVID-19-Like Epidemics: Methods and Challenges. *IEEE Network*, *35*(3), 27-33.

Lepri, G., Orlandi, M., Lazzeri, C., Bruni, C., Hughes, M., Bonizzoli, M., & Matucci-Cerinic, M. (2020). The emerging role of lung ultrasound in COVID-19 pneumonia. *European journal of rheumatology*, *7*(Suppl 2), S129.

Eccles, R., Fietze, I., & Rose, U.B. (2014). Rationale for treatment of common cold and flu with multi-ingredient combination products for multi-symptom relief in adults. *Open Journal of Respiratory Diseases*, *4*(03), 73.

Tuli, S., Tuli, S., Tuli, R., & Gill, S.S. (2020). Predicting the growth and trend of COVID-19 pandemic using machine learning and cloud computing. *Internet of Things*, *11*.

Malki, Z., Atlam, E.S., Ewis, A., Dagnew, G., Alzighaibi, A.R., ELmarhomy, G., & Gad, I. (2021). ARIMA models for predicting the end of COVID-19 pandemic and the risk of second rebound. *Neural Computing and Applications*, *33*(7), 2929-2948.

Doroshenko, A. (2020). Analysis of the distribution of COVID-19 in italy using clustering algorithms. *In IEEE Third International Conference on Data Stream Mining & Processing (DSMP)*, 325-328.

Gupta, V.K., Gupta, A., Kumar, D., & Sardana, A. (2021). Prediction of COVID-19 confirmed, death, and cured cases in India using random forest model. *Big Data Mining and Analytics*, *4*(2), 116-123.

Bhati, A., & Jagetiya, A. (2020). Prediction of covid-19 outbreak in india adopting bhilwara model of containment. *In 5th International Conference on Communication and Electronics Systems (ICCES)*, 951-956.

Kurniawan, R., Abdullah, S.N.H.S., Lestari, F., Nazri, M.Z.A., Mujahidin, A., & Adnan, N. (2020). Clustering and Correlation Methods for Predicting Coronavirus COVID-19 Risk Analysis in Pandemic Countries. *In 8th International Conference on Cyber and IT Service Management (CITSM)*, 1-5.

Ibrahim, A.M., Eid, M.M., Mostafa, N.N., Bishady, N.E.H.M., & Elghalban, S.H. (2020). Modeling the effect of population density on controlling Covid-19 initial Spread with the use of MATLAB numerical methods and stringency index model. *In 2nd Novel Intelligent and Leading Emerging Sciences Conference (NILES)*, 612-617.

Darapaneni, N., Reddy, D., Paduri, A.R., Acharya, P., & Nithin, H.S. (2020). Forecasting of COVID-19 in India using ARIMA model. *In 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, 0894-0899.

https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset

Vijayakumar, J., & Arumugam, S. (2013). Certain investigations on foot rot disease for betelvine plants using digital imaging technique. *In International Conference on Emerging Trends in Communication, Control, Signal Processing and Computing Applications (C2SPCA)*, 1-4.

Kandasamy, R., & Krishnan, S. (2015). Enhanced energy efficient method for WSN to prevent far-zone. *International Journal on Communications Antenna and Propagation (IRECAP)*, *4*(4), 137-142.

Abdekhoda, M., Sattari, A., Mohammadi, M., & Salih, K.M. (2020). Presenting a conceptual model of adopting micro-blogging in learning. *Webology*, *17*(1), 98-108.