**PREDICTION AND DISTRIBUTION OF COVID-19 USING PROPHET, K-MEANS AND HIERARCHICAL CLUSTERING ALGORITHM**

1. **Mr.R.VINSTON RAJA M.Tech., [2] S.PAUL TIMOTHY [3] P.V.ARAVINTH [4] J.HARIPRASATH**

# Assistant Professor [2]UgScholar [3]UgScholar [4]UgScholar [1][2][3][4] Department Of Information Technology

**[1][2][3][4] Panimalar Institute Of Technology [1]**[winstonroja@gmail.com](mailto:winstonroja@gmail.com)**[2]**[paultimothysrs@gmail.com](mailto:suryaprakash@gmail.com)**[3]**[aravinth0399@gmail.com](mailto:arasu@gmail.com)**[4]**[harihvj2000@gmail.com](mailto:shurihardy@gmail.com)

# Abstract

**Coronaviruses are a group of viruses that cause various diseases in mammals and birds. In humans, they cause a range of respiratory disorders. The epidemic of coronavirus disease-2019 (COVID19) establishes a medical emergency of worldwide concern with an exceptionally high danger of spread and affect the entire world. New cases were recorded in almost all countries. However, in some countries, the level of the disease remained low, while in other countries, they became “red zones” - with a high number of sick and dead. In this project, by using Data Science algorithm- K-Means and Hierarchical Clustering algorithm each country will be divided into groups or segments based on the COVID patients count. In this project, Clustering models has been created across the countries in the world and across the states in India, and the performance of the model is compared. This project also forecasts the future COVID count for India. By using this result, set of countries which is having higher COVID count can be easily visualized and the appropriate actions will be taken to decrease the count**

**Keywords-Between sum of squares (BSS),Total sum of squares(TSS),Within sum of squares(WSS).**

# 1.INTRODUCTION

# Coronavirus disease (COVID-19) is an infectious disease caused by a newly discovered coronavirus. The world nowadays faces the COVID-19 pandemic, which has a very large global spread that importantly leads to a large number of deaths around the world. Forecasting the spread of coronavirus and finding the distribution of COVID cases across the countries is one of the challenges in recent times. Forecasting the pandemic with high accuracy will help different countries prepare a plan to fight a war against the virus spread.

# Data Science techniques are extensively used for modelling real-world problems. Specifically, to predict diseases have been used extensively in recent times. The important trend in Information Technology is to identify the meaningful information from the enormous amount of data stored in files, databases, and other repositories and to develop powerful means for analysis and interpretation of such data for the extraction of interesting knowledge that could help in decision-making

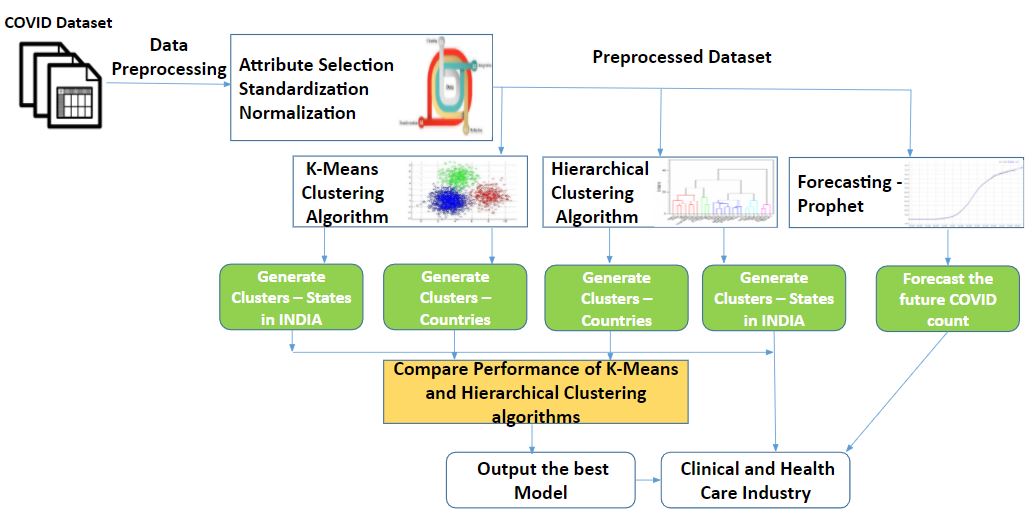
# PROPOSED WORK

Data Science techniques of K-Means and Hierarchical clustering have been applied to form the clusters among countries and states. It can handle both categorical and continuous data. Data Science is an interdisciplinary field that incorporates computer science, mathematics, statistics and domain knowledge. Using data science algorithms, convergence is guaranteed. Clustering for the world – grouping countries in the world according to the COVID count and Clustering for India – Grouping states in India according to the COVID count. Using Prophet, forecasting on time series data is performed, to predict the future COVID count of India.

# METHEDOLOGY

1. **IMPLEMENTATION:**

In Data importing, input file for doing the clusters among the countries in the world, is a csv file taken from world health organization dataset. The input file for doing the clusters among the states in India is also a csv file taken from covid19.india dataset. Once data have been collected, it have been imported into the R environment for clustering and prediction. 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 and form the groups among the states in India using COVID patient’s count. Hierarchical clustering algorithm formed the clusters among the countries and states among the India, using the COVID patient’s count. 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.

****

**System architecture**

**B.MODULES**

There are five Modules

➢ Importing COVID Dataset and Data Preprocessing

➢ K-Means Algorithm – Clustering Countries and States

➢ Hierarchical Algorithm -Clustering Countries and States

➢ Comparison of K- Means and Hierarchical Clustering Algorithms

➢ Forecasting future COVID count for India Using Prophet

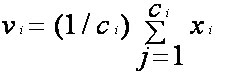
**C. DATA IMPORTING AND PREPROCESSING**

Data is available in any file format like .txt, .csv, .xlsx, .spss etc. Data have to be loaded in to R environment for analysis. Once data have been extracted from the file it should be stored in a data frame. Libraries necessary for forecasting have to be loaded into the R environment. For forecasting the future values in the time series data, prophet package has to installed and loaded into the program. Install some more packages for subsequent operations like, dplyr package used for applying the sql query to extract the India dataset from the world, ggplot2 package to give some more visualization to the output and lubridate package to convert the datestamp in the format of YYYYMMDD. Data pre-processing is the data mining technique that involves transforming raw data into understandable format. The Raw data is highly susceptible to noise, missing values, and inconsistency. Real world data is often incomplete, inconsistent, and is likely to contain errors. The missing Values problem have to be solved by simple statistical techniques.12 Data preprocessing is the proven method for resolving such issues. In order to improve the quality of the data consequently, the mining results of raw data is pre-processed so the efficiency process improved. It is not necessary to hold all the attributes for doing the analysis, we can hold only the attributes which is affecting the analysis. Data standardization is the process by which similar data is collected in various formats is transformed to a common format that enhances the comparison process, allows for collaborative research and largescale analytics. Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

**D.K-MEANS ALGORITHM – CLUSTERING COUNTRIES AND STATES**

Clustering is one of the most common exploratory data analysis technique used to get an intuition about the structure of the data. It can be defined as the task of identifying subgroups in the data such that data points in the same subgroup are very similar while data points in different clusters are very different. The process of organizing objects into groups whose members are similar in some way is known as clustering. K Means algorithm is an iterative algorithm that tries to partition the dataset into Kpre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. The group of similar objects form clusters and dissimilar objects belong to another cluster. Fast, robust and easier to understand.

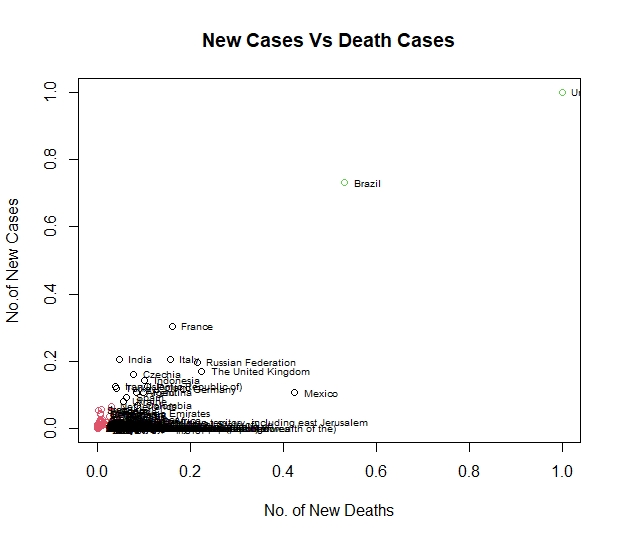
Let X = {x1,x2,x3,……..,xn} be the set of data points and V = {v1,v2,…….,vc} be the set of centres. 1) Randomly select ‘c’ cluster centres. 2) Calculate the distance between each data point and cluster centres. 3) Assign the data point to the cluster centre whose distance from the cluster centre is minimum of all the cluster centres. 4) Recalculate the new cluster centre using: where, ‘ci’ represents the number of data points in i th cluster.



5) Recalculate the distance between each data point and new obtained cluster centres. 6) If no data point was reassigned then stop, otherwise repeat from step 3.

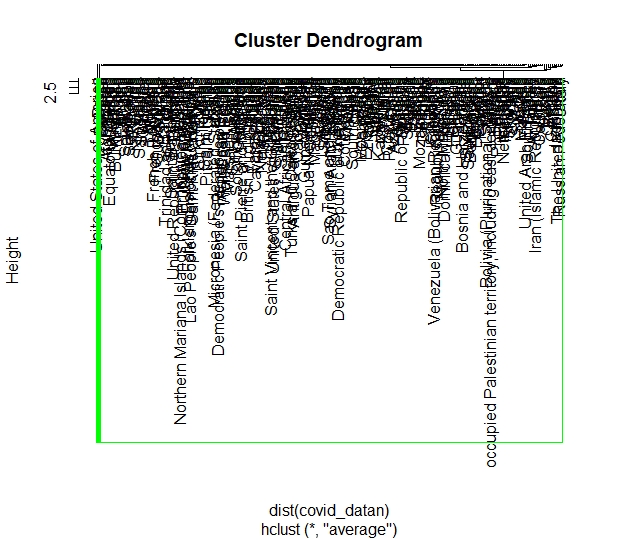
Distance between the data items is measured by Euclidean distance method.

 d=|x-y|=sqrt(sum_(i=1)^n|x_i-y_i|^2). 



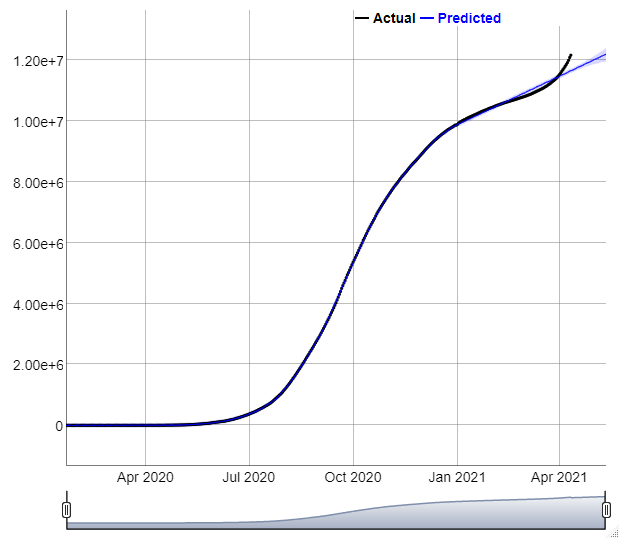
**E. HIERARCHICAL ALGORITHM – CLUSTERING COUNTRIES AND STATES**

A Hierarchical clustering method works via grouping data into a tree of clusters. Hierarchical clustering begins by treating every data point as a separate cluster. In Hierarchical Clustering, the aim is to produce a hierarchical series of nested clusters. A diagram called Dendrogram (A Dendrogram is a tree-like diagram that statistics the sequences of merges or splits) graphically represents this hierarchy and is an inverted tree that describes the order in which factors are merged. Initially consider every data point as an individual Cluster and at every step, merge the nearest pairs of the cluster. At first every data set is considered as individual entity or cluster. At every iteration, the clusters merge with different clusters until one cluster is formed.16 Algorithmic steps for Hierarchical clustering algorithm: 1. Start by assigning each item to its own cluster, so that if N items are there, then N clusters will be there, each containing just one item. Let the distances (similarities) between the clusters equal the distances (similarities) between the items they contain. 2. Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one less cluster. 3. Compute distances (similarities) between the new cluster and each of the old clusters. 4. Repeat steps 2 and 3 until all items are clustered into a single cluster of size N Let us take six objects (A,B,C,D,E,F). Each object have two measured features(X1,X2). Assign each item to its own clusters. Thus, in the beginning there are 6 clusters. Distance matrix: Use the Euclidean Distance formula to compute the distance between the objects and form the distance matrix. dist((x, y), (a, b)) = √(x - a)² + (y - b)²

****

**F. FORECASTING FUTURE COVID COUNT FOR INDIA USING PROPHET**

Time Series Analysis is a way of analyzing and learning the behavior of datasets over a period. Moreover, it helps in learning the behavior of the dataset by plotting the time series object on the graph. Prophet is open source software released by Facebook’s Core Data Science team. Prophet is a procedure for forecasting time series data where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects.22 It works best with time series that have strong seasonal effects. Prophet is robust to missing data and shifts in the trend, and typically handles outliers well. It is available for download on CRAN and PyPI. It helps businesses to learn the behavior of their products by forecasting prices, sales, or weather. It uses a decomposable time series model with three main model components: trend, seasonality, and holidays. y(t) = g(t) + s(t) + h(t) + €t y(t) - Forecast g(t) - trend function s(t) - represents periodic changes (e.g., weekly and yearly seasonality) h(t) - effects of holidays which occur on potentially irregular schedules εt - error term accounts for any unusual changes not accommodated by the model A trend is a shift in development either in increment or decrement direction. Seasonality is a feature of time series object that occurs at a particular time/season and changes the trend. Holidays are a time period that changes a lot to the business. It can make a profit or loss depending upon the business. Prophet is used in many applications across Facebook for producing reliable forecasts for planning and goal setting. Prophet performs better than any other approach in the majority of cases. Prophet provide a reasonable forecast on messy data with no manual effort. Prophet is robust to outliers, missing data, and dramatic changes in your time series. The Prophet procedure includes many possibilities for users to tweak and adjust forecasts



**H. CONCLUSION AND FUTURE ENHANCEMENT**

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. This project concludes 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 project 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 project, for 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.

**I. Reference**

1. Anastasiya Doroshenko., "Analysis of the Distribution of COVID-19 in Italy Using Clustering Algorithms", IEEE International Conference on Data Stream Mining & Processing, pp:325-328, 2020

2. Ginalber L.O.Serra, Daiana Gomes, "Machine Learning Model For Computational Tracking and Forecasting the COVID-19 Dynamic Propagation”,IEEE Journal of Biomedical and Health Informatics, 2021.

3. Vishan Kumar Gupta, et.a., " Prediction of COVID-19 confirmed, death, and cured cases in India using random forest model", IEEE Big Data Mining and Analytics - Volume: 4, Issue: 2, June 2021.

4. Rahmad Kurniawan et.al., " Clustering and Correlation Methods for Predicting Coronavirus COVID-19 Risk Analysis in Pandemic Countries", IEEE International Conference on Cyber and IT Service Management, 2020.

5. Valerio Bellandi; Paolo Ceravolo; Samira Maghool; Stefano Siccardi, “A Comparative Study of Clustering Techniques Applied on Covid-19 Scientific Literature” IEEE International Conference on Internet of Things: Systems, Management and Security, 2020.

6. C. Huang et al., "Clinical features of patients infected with 2019 novel coronavirus in Wuhan China", Lancet, vol. 395, no. 10223, pp. 497-506, 2020.64

7. Remigio Ismael Hurtado Ortiz; Juan Carlos Barrera; Katherine Michelle Barrera., "Analysis model of the most important factors in Covid-19 through data mining, descriptive statistics and random forest", IEEE International Autumn Meeting on Power, Electronics and Computing, pp.4-6, 2020..

8. N. Zhu et al., "A novel Coronavirus from patients with Pneumonia in China 2019", N. Engl. J. Med., vol. 382, no. 8, pp. 727-733, 2020.

9. Pınar Cihan, "Fuzzy Rule-Based System for Predicting Daily Case in COVID-19 Outbreak", IEEE International Conference on Multidisciplinary Studies and Innovative Technologies, 2020.

10. S. S. Chenar and Z. Deng, "Development of genetic programming-based model for predicting oyster norovirus outbreak risks", Water research, vol. 128, pp. 20-37, 2018