Knowledge-based approach for predicting covid-19 using Machine learning

Subradev Sarkar[1], Suva Ghosh[2], Tanmoy Paul[3], Sourav Das[4], Dharmpal Singh[5], Sudipta Sahana[6]

[1 2 3 4] B. Tech, Department of Computer Science and Engineering, JIS College of Engineering, Kalyani, West Bengal, India

[5]Head of the Department, Department of Computer Science and Engineering, JIS College of Engineering, Kalyani, West Bengal, India

[6]Assistant Professor, Department of Computer Science and Engineering, JIS College of Engineering, Kalyani, West Bengal, India

Abstract - **For the last two years, the entire world is suffering from covid-19 diseases. It declares a global pandemic by the WHO. At the earlier stages, Indian people were not aware of this virus and now the cases are increasing. The technique proposed in this paper is an efficient solution for identifying the covid cases and will help us to know whether a person is covid positive or not without medical testing. The technology behind this project is Machine Learning. Which is an advanced and efficient solution to make predictions easy and give a perfect result. In this project, our effort is that based on the symptoms of the covid-19 we can predict whether a person is suffering from covid-19 or not. This method is possible by using patients' health reports where it has been recorded that, which patent is covid positive and which one is negative and can help create knowledge-based by using Machine learning. So, after collecting the data first data-cleaning processes have been done. Then after data-cleaning, correlations have been found to identify the most important features by using the Factor analysis technique. From Factor analysis, the Total effect value has been collected, then the Total effect value has been divided into two clusters 0 and 1. After that based on the cluster-centers a knowledge-based have been created, From where it can be determined that if the total effect value of the new test data is in the range of any cluster value, it will belong to that particular cluster and can be easily said that if it belongs to 0 clusters the result will be negative or if it belongs to 1 cluster the result will be positive.**

Index Terms - **Covid-19, Healthcare, Machine Learning, Factor analysis, Clustering, K-means, Hierarchical, Knowledge-based.**

I.INTRODUCTION

Coronavirus disease 2019 (COVID-19) is caused by a new virus SARS-Cov-2 first identified in Wuhan, China, in December 2019. At the beginning of this disease, symptoms were not that much clear as now but there are still new strains being found in patients. So from all the symptoms, we choose the most affected symptoms, like: Breathing Problem, Fever, Dry Cough, Sore throat, Running nose, Asthma, Heart Disease, whether the person has traveled Abroad or Contact with COVID Patient, Attended Large Gathering, Visited Public Exposed Places and is there any one of the Family working in Public Exposed Places, etc. Our project is based on text data where we have several patient data having both covid positive and negative, this data helped us to figure out the important factors. We have collected our data from the Kaggle website according to our needs. After that, we have used machine learning technology to create a knowledge-based from which we can then easily predict whether a person having such symptoms is covid-19 positive or not.

II. LITERATURE SURVEY

Machine learning (ML), Deep learning (DL), Natural Language Processing (NLP) are the most used technologies that are used in different fields like IT industries, Heal sector, etc. For this data play role. A larger amount of data is required for Machine Learning and Natural Language Processing to create models for pattern recognition, analysis, and prediction. In recent time's text analytics, text mining has gained much interest because of NLP, and also various algorithms are used for this, like Classification for text mining. Kumar et al [1] performed a Swaot analysis of various supervised and unsupervised text classification algorithms for mining the unstructured data. The different examples where text classification applies are sentiment analysis, fraud detection, spam detection, etc. Machine learning has changed the way of diagnosis by giving a great result in various diseases like diabetes, heart disease prediction, etc. Sarwar et al [2] diagnose diabetes by using ML and ensemble learning techniques. Mustafa S.Kadhm et al [3] use k-means clustering and classification methods to predict Diabetes the proposed system achieves a higher classification rate than the other system. Purvashi Mahajan et al [4] proposed a clustering algorithm in the proposed paper to find out how the k-means algorithm works in Disease prediction and figure out that k-means can be used for these kinds of prediction but it should be used with another algorithm for better result. Fahad Ahmad et al. [5] performed Prediction of COVID-19 Cases Using ML for Effective Public Health Management. The outcome of this study points that the human development index and population density can also be associated with the number of COVID-19 cases in an area. Different techniques have been used like Spearman’s Rank Correlation, Shallow Single-Layer Perceptron Neural Networks (SSLPNN), and Prediction of COVID-19 Cases by Regression Analysis. Khanday et al. [6] did a machine learning-based approach for detection COVID-19 using clinical text data. Feature engineering is used in their project for extracting various features as per the semantics and converting them into probabilistic values. Traditional machine-learning algorithms like Logistics Regression, Naive Bayes, Support Vector Machine, and Decision Trees are used for best model fit and Logistic regression and naive Bayes classifiers give a good amount of accuracy of 96.2%. As ML is used in the diagnosis and prediction of diseases so for COVID-19 it is also a useful technique. According to various research, less work has been done on diagnosis and prediction using text data, so we used machine learning to predict COVID-19 cases.

III. METHODOLOGY

MACHINE LEARNING

Machine learning is a subset of artificial intelligence. Artificial intelligence is an intelligence system almost like human intelligence but works according to some algorithms. It is a technique that enables machine learning to mimic human behaviour. As Machine Learning is a part of AI, so it also needs some algorithms to work with and use statistical methods to enable machines to improve with experience. For ML first, it needs some data to recognize the pattern, and then from the study of pattern recognition, machine learning explores the study and builds algorithms that can learn from and make predictions on data.

SUPERVISED LEARNING

Supervised learning is a part of machine learning, where it is used labelled data for training. The data for training consists of a set of training examples. Each example is a pair consisting of the input object and the desired output value. A supervised learning algorithm analyses the training data and predicts output based on the training dataset. In supervised learning output is given with the training data, the machine just has to recognize the pattern of the training data and make a prediction according to the pattern for the new data.

UNSUPERVISED LEARNING

Unsupervised learning is another part of machine learning where it determines a function to describe hidden structures from unlabelled data. The machine is trained on unlabelled data without any guidance. As the trained data given to the learner are unlabelled so it is not possible to find out the accuracy of the structure by trained model output.

Unsupervised learning can be used to solve association problems and clustering problems. Association problems involve discovering patterns in data, finding co-occurrences, and so on. Clustering problems involves discovering similarities between data and then divided them into separate groups, then these groups are called clusters.

ML-Algorithm Clustering:

Clustering is the process of dividing the datasets into groups, consisting of similar data points. Clustering methods are used to find similarities as well as the relationship among data and then divide the data into groups having similarities based on features. First, it finds out the common points among data and then tries to separate the data according to the data point. When a data is closer to the data point then the clustering methods take them and make a separate group.

K-means:
K-means clustering is a kind of Exclusive clustering, here Data-points belong exclusively to one cluster and its main goal is to group similar data points into a cluster. The number of groups or clusters is represented by K. The algorithm then runs iteratively to assign each data-point to one of the K groups based on the features that are already pointed.

K-means clustering algorithm requires two inputs:
i) k = number of cluster.
ii) Training dataset.

Working of K-means algorithm:

- Step 1 - First, we need to specify the number of clusters, K. In case it is not predetermined then we have to apply elbow methods to find out the optimal number of the clusters.
- Step 2 - Next, randomly initialize K points called cluster centroids. The value of k can be determined by the elbow curve.
- Step 3 - Now it will compute the distance between the data points and the cluster centroid initialized. Depending upon the minimum distance, data points are divided into groups.
- Step 4 - Compute the mean of clusters. Then reposition the cluster centroid to the mean.
- Step 5 - Repeat the previous two steps iteratively till the cluster centroids stop changing their positions.
- Step 6 – finally, the k-means clustering algorithm converges and divides the data points into desire clusters.

Hierarchical:
Hierarchical clustering or hierarchical cluster analysis or HCA is a method of clustering that seeks to build a hierarchy of clusters in a given dataset. Hierarchical clustering generally falls into two types:

- Agglomerative: This is a "bottom-up" approach: Initially, each point is a cluster and repeatedly combines the two “nearest” clusters into one.
- Divisive: This is a "top-down" approach: Start with one cluster and recursively split the cluster.

Agglomerative Hierarchical clustering:
The algorithm initially takes each data point as a single cluster and then starts forming clusters according to the distance between each data. This process continues until all the clusters form a single cluster that contains all the datasets.
Working of Agglomerative Hierarchical clustering:

- Step 1 – Initially each data point forms a cluster.
- Step 2 – Compute the distance matrix between the clusters.
- Step 3 – merge the two closest clusters and update the distance matrix.
- Step 4 - Repeat Step 3 until only a single cluster remains.
- Step 5 – If cluster numbers are not predefined to perform the hierarchical clustering, then dendrogram is the most effective way to find out the cluster number.

IV. IMPLEMENTATION

The proposed methodology consists of 4.1 to 4.7 steps. To implement the project first we have collected the data. Then we have performed some data cleaning processes to refine the data. After that factor analysis has applied to reduce the dimension of the dataset and divided them into two clusters. The visual representation of the process (Figure 1) is shown below.

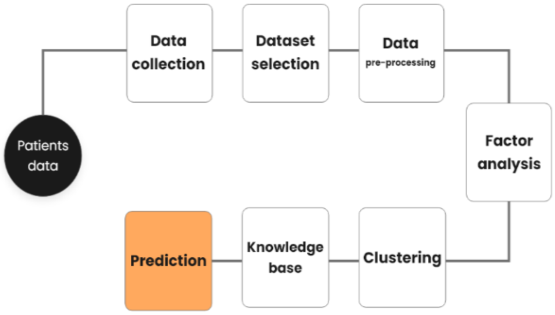

Figure 1: Process Flowchart

4.1 Data collection: As Covid-19 has become a global pandemic and WHO declared this as a Health Emergency. Hospitals and researchers give open access to the data regarding this pandemic. One of the open-source data resources is Kaggle, where many data are openly found and also many research works are found. We have collected data from Kaggle. The dataset we have selected is a symptom-based dataset along with the covid result. The dataset has 21 attributes and 5435 patient data.

4.2 Data selection: Our data has been collected from Kaggle. We have collected many data but from all of them, we have only chosen the relevant dataset for our further work process. Our dataset consists of 5435 patients data and 21 attributes namely Breathing Problem, Fever, Dry Cough, Sore throat, Running nose, Asthma, Chronic Lung Disease, Headache, Heart Disease, Diabetes, Hypertension, Fatigue, Gastrointestinal, Abroad travel, Contact with COVID Patient, Attended Large Gathering, Visited Public Exposed Places, Family working in Public Exposed Places, Wearing Masks, Sanitization from Market, COVID-19.

4.3 Data pre-processing: In data pre-processing, first we have checked null values then checked if there is any unique value present or not. In our dataset, we have 21 attributes but there were no null values present in our dataset. But in the unique values check, we found that two attributes have only one type of value so we eliminate those attribute columns “Wearing Masks” and “Sanitization from Market”.

4.4 Factor analysis: Our dataset has many attributes so we have used factor analysis to find out the important factors. For doing factor analysis first we have found a correlation between attributes. Then figure out the Eigenvalue and Eigenvector. After found Eigenvalues and Eigenvectors we have calculated the percentage contribution of Eigenvalues to know which factor is most important and which one is less and according to that result, we ignore the less important factor. But in ours, every attribute or factor has a kind of equal contribution so we keep all the factors (Table 1). Percentage contribution of Eigenvalues = (Eigenvalue/ sum of all Eigenvalue)*100

Table 1: Percentage contribution of Eigenvalues

Factors	Eigenvalue	Percentage
Breathing Problem	2.290287811	12.72382117
Fever	0.468431262	3.602395901
Dry Cough	1.44048088	8.002671555
Sore throat	1.271069585	7.061497697
Running nose	1.261876792	7.010426621
Asthma	0.602303823	3.346132348
Chronic Lung Disease	0.635278465	3.529324803
Headache	0.68197592	3.788755109
Heart Disease	1.168848334	6.493601855
Diabetes	0.740954107	4.116411708
Hyper Tension	0.792220099	4.401222772
Fatigue	0.813223714	4.517909525
Gastrointestinal	0.869898556	4.832769757
Abroad travel	0.909205526	5.05114181
Contact with COVID Patient	0.945878305	5.254879473
Attended Large Gathering	0.995750316	5.531946201
Visited Public Exposed Places	1.067102635	5.92834797
Family working in Public Exposed Places	1.04521387	5.806743724

After finding the Percentage contribution of Eigenvalues we have calculated the Contribution of each Eigen Vector Corresponding to each Eigen Value (Table 2)

Contribution of Eigen Vector Corresponding Eigen Value = (√Eigenvalue × square (Eigenvector))

Table 2: Eigen Vector Corresponding Eigen Value

	Breathing Problem	. . .	Family working in . . .	sum
Eigenvalue	2.290288	. . .	1.045214	0.96981877
Breathing Problem	0.215065	. . .	0.001572	0.9576762
Fever	0.249484	. . .	0.001201	0.97032017
Dry Cough	0.206539	. . .	0.003411	0.99517771
Running Nose	0.004701	. . .	0.237635	0.96981877

Using the sum from Table 2 the Total effect value has been calculated (Table 3).

Total effect value = (0.96981877 * Breathing Problem + 0.9576762 * Fever + 0.99081731 * Asthma+ …..)

Table 3: Total effect value

Breathing Problem	Fever	. . .	Visited Public	total effect value
1	1	. . .	1	11770286.98
1	1	. . .	1	9786979.073
1	1	. . .	0	11771007.95
1	1	. . .	1	11816685.25

4.5 Total Effect value: From factor analysis, we have got Total effect value (Table 3). Then the total effect value has been divided into two clusters 0 and 1 for a better understanding and value evolution.

4.6 Clustering: As the Total effect value is a kind of continuous type of value, so it is difficult to predict cases, as we are expected to result in 0 or 1, which means negative or positive. For getting this result we have divided our data into two clusters 0 and 1 and for this clustering, we have used the K-means (Table 4) and Hierarchical (Table 5) clustering algorithm and took cluster number as 2.

Table 4: K-means Cluster data

Sl no.	total effect value	Cluster center	cluster
2171	6.839775e+06	7.392814e+06	0
2859	6.825443e+06	7.392814e+06	0
2860	8.816203e+06	7.392814e+06	0
…	…	…	…
1106	1.373891e+07	1.126094e+07	1
1105	1.274845e+07	1.126094e+07	1
1104	1.177771e+07	1.126094e+07	1

Table 5: Hierarchical Cluster data

Sl no.	total effect value	Cluster center	cluster
2171	6.839775e+06	8236924	0
2819	9.785526e+06	8236924	0
2820	8.800908e+06	8236924	0
…	…	…	…
1231	1.173849e+07	11874725	1
2748	1.079587e+07	11874725	1
2749	1.475222e+07	11874725	1

4.7 Knowledge-based: From total effect values we have got 2 cluster centers and then link them with their corresponding cluster and sort the final result according to cluster value. After that, we have to find out the max and min range of every cluster's total effect value. From this knowledge-based, we can predict the output that if the total effect value of new test data is in between any closet range of 0 or 1's cluster centers value we can easily tell that the person is coved positive or negative.

After applying:
K-means clustering on "total effect value" we have got two cluster centre (Table 5) for
0 -> cluster centre is 7.392814e+06; and for 1-> cluster centre is 1.126094e+07.

Hierarchical clustering on "total effect value" we have got two cluster centre (Table 6) for
0 -> cluster centre is 8.236924e+06; and for 1-> cluster centre is 1.187473e+07.

Knowledge-based: According to the cluster center
K-means:
For 0 cluster ->
Min value of "total effect value" is: 2.948229e+06
Max value of "total effect value" is: 8.918220e+06
For 1cluster->
Min value of "total effect value" is: 9.764823e+06
Max value of "total effect value" is: 1.475482e+07
Hierarchical:
For 0 cluster ->
Min value of "total effect value" is: 2.948229e+06
Max value of "total effect value" is: 9.899007e+06
For 1cluster->
Min value of "total effect value" is: 1.075058e+07
Max value of "total effect value" is: 1.475482e+07

Table 6: K-means Knowledge-based creation

Sl no.	K-means			cluster
	total effect min	total effect max	Cluster center	
1	2.948229e+06	8.918220e+06	7.392814e+06	0
2	9.764823e+06	1.475482e+07	1.126094e+07	1

Table 7: Hierarchical Knowledge-based creation

Sl no.	Hierarchical			cluster
	total effect min	total effect max	Cluster center	
1	2.948229e+06	9.899007e+06	8.236924e+06	0
2	1.075058e+07	1.475482e+07	1.187473e+07	1

From this (Table 7 & Table 6) Knowledge-based we can say that if the total value effect of the test data is in the range of "total effect min" to "total effect max" it will belong to that particular cluster. Ex: If it belongs to 0 clusters then the result will be negative.

V. RESULTS AND DISCUSSION

5.1 Test data:

We have taken around 10% of the data from our main dataset for testing before creating the knowledge base. Then apply the Factor analysis technique to find out the total effect value and based on that we have predicted the result according to our created knowledge-based.

5.2 Prediction:
From our test data for
K-means prediction, those "Test data total effect values" are in between the range of 2.948229e+06 to 8.918220e+06 that belongs to cluster 0 and the result is negative. Again those "Test data total effect value" are in between the range of 9.764823e+06 to 1.475482e+07 that belongs to cluster 1 and the result is positive.

Hierarchical prediction those "Test data total effect values" are in between the range of 2.948229e+06 to 9.899007e+06 that belongs to cluster 0 and the result is negative. And those "Test data total effect value" are in between the range of 1.075058e+07 to 1.475482e+07 that belongs to cluster 1 and the result is positive according to.

5.3 Result:
Accuracy-Score:
K-means: 71% accuracy
Hierarchical: 60% accuracy
According to the Accuracy-score K-means clustering method is best to predict the result for our dataset.

VI. CONCLUSION AND FUTURE SCOPE

In this project, we have studied many cases and done various Visualization based on data and there are still some things left to do. We have done data pre-processing and factor analysis on our dataset. From FA (factor analysis) we have got the total effect value and then divided it into clusters. These clusters and the cluster centers are then used for knowledge base creation that easily predicts the test result. And according to the result, K-means is the best method for our project compare to Hierarchical although the accuracy is low. As the dataset, we have chosen for this project has a maximum of two clusters there, so for this reason, we are getting some deflection and for this reason the accuracy–score is not that much well.

ACKNOWLEDGMENT

The study and the analysis of the project wish to express our gratefulness to Dr. Dharmpal Singh and Dr. Sutipta Sahana for giving us the golden opportunity to do this wonderful project on the topic "PREDICTING COVID 19 CASES USING MACHINE LEARNING" and also giving us effective guidance in the development of this project. Their conspiration of the subject and all the helpful hints, they provided to the successful development of this project, which also helps us in doing a lot of research and come to know so many new things.
We also sincerely thankful to Prof. (Dr.) Partha Sarkar Principal of JIS College of Engineering for allowing us to carry out the project work.

REFERENCES

[1] Akshi Kumar, Vikrant Dabas, Parul Hooda (2018) "Text classification algorithms for mining unstructured data: a SWOT analysis". Int J Inf Technol.

[2] Abid Sarwar, Mehbob Ali, Jatinder Manhas & Vinod Sharma (2018) "Diagnosis of diabetes type-II using hybrid machine learning-based ensemble model". Int J Inf Technol.

[3] Mustafa S. Kadhm, Ikhlas Watan Ghindawi, Duaa Enteesha Mhawi "An Accurate Diabetes Prediction System Based on K-means Clustering and Proposed Classification Approach" Research ISSN 0973-4562 Volume 13, Number 6 (2018) pp. 4038-4041.

[4] Purvashi Mahajan, Abhishek Sharma "Role of K-Means Algorithm in Disease Prediction" ISSN: 2319-7242 Volume 5 Issue 4 April 2016, Page No. 16216-16217

[5] Fahad Ahmad, Saleh N. Almuayqil, Mamoona Humayun, Shahid Naseem, Wasim Ahmad Khan, Kashaf Junaid (2021) "Prediction of COVID-19 Cases Using Machine Learning for Effective Public Health Management".

[6] Akib Mohi Ud Din Khanday, Syed Tanzeel Rabani, Qamar Rayees Khan, Nusrat Rouf and Masarat Mohi Ud Din (**2020**) "Machine learning-based approaches for detecting COVID-19 using clinical text data".