Hybrid of K-Means and partitioning around medoids for predicting COVID-19 cases: Iraq case study

Nidaa Ghalib Ali[1], Saba Dhey Abed[2], Faris Ali Jasim Shaban[3], Korakod Tongkachok[4], Samrat Ray[5], Refed Adnan Jaleel[6]

[1] Technical Computer System, Al-Furat Al-Awsat Technical University, Iraq
[2] Computer Science, Baghdad Institute of Technology, Iraq
[3] Electronic Engineering, Alnukhba University College, Iraq
[4] Department: Faculty of Law, Thaksin University, Thailand
[5] The Institute of Industrial Management, Economics and Trade Peter The Great Saint Petersburg Polytechnic University, Russia
[6] Information and Communication Engineering, Al-Nahrain University, Iraq

ABSTRACT

COVID-19 was discovered near the end of 2019 in Wuhan, China. In a short period, the virus had spread throughout the entire world. One of the primary concerns of managers and decision-makers in all types of hospitals nowadays is to implement detection plans for status of patient (Negative, Positive) in order to provide enough care at the proper moment. To reduce a pandemic of COVID-19, improving health care quality could be advantageous. Making clusters of patients with similar features and symptoms supplies an overview of health quality given to similar patients. In the scope of medical machine learning, the K-means and Partitioning Around Medoids (PAM) clustering algorithms are usually used to produce clusters depend on similarity and to detect helpful patterns from sizes of data. In this paper, we proposed a hybrid algorithm of K-Means and Partitioning Around Medoids (PAM) called K-MP to take benefits of both PAM and K-Means to construct an efficient model for predicting patient status. The suggested model for the real dataset was collected from 400 patients in the many Iraqi clinics using a questionnaire. We evaluated the proposed K-MP by using true negative rate, balance accuracy, precision, accuracy, recall, mean absolute error, F1 score, and root mean square error. From these performance measures, we found that K-MP is more efficient in discovering patient status comparing to K-Means and PAM.

Keywords: Unsupervised medical machine learning, Clustering, K-MP, Evaluation, Diagnostic

Corresponding Author:

Refed Adnan Jaleel
Information and Communication Engineering Department
Al-Nahrain University, Iraq
Email: iraq_it_2010@yahoo.com

1.   Introduction

In most of the world's countries, COVID-19 is like a storm to the world. The entire world is attempting to cope with this pandemic. COVID-19 is spread through contact, sneezing, coughing, or talking. Thus, the rapid expansion of COVID-19 is a risk. A country's ability to combat the pandemic may be aided by improved healthcare quality. In order to make things worse, up to 72 hours of COVID-19 survive on areas that are apparently normal, making individuals get it [1].

Early detection is the greatest method to enhance the COVID-19 survival rate and medical data mining tools can be used in the work now under progress. In the medical domain, medical machine learning is widely applied. It is commonly believed that medical machine learning can tackle the pressing problems linked to COVID-19 if there is patterns information gathered about the patients [2]. Medical machine learning is applied to allow doctors to make accurate process of diagnostic and deduce diagnostic rules.

In the scope of data science and medical machine learning, both medical unsupervised and medical supervised learning are commonly used to solve different types of real-world problems [3].

Creating clusters of patients with similar symptoms of COVID-19 quality gives help to vision into the quality of care provided to diverse patients [4]. Unsupervised medical algorithms of machine learning are being used

in the case of clustering. Typically, unsupervised medical machine learning algorithms distinguish vision construction of the data from unlabeled data consisting in the data set [5] [6].

Then, based on the similarity of the hidden dataset constructions, the clustering algorithms divide and find the data points. K-Means and PAM (K-Medoids) clustering are the most useful and widely applied to automatically distinguish the hidden structure [7].

Based on patient data, this paper offers predicting corona possibilities (negative, positive) using the three algorithms of unsupervised medical machine learning (K-Means, PAM, and K-MP). The prediction model is assessed using the confusion matrix for the performance of the 400 COVID-19 patient data set. This paper looks at the classification reports for the K-Means, PAM, and K-MP algorithms. The findings of this paper offer that the proposed K-MP algorithm findings more effectively for COVID-19 dataset.

This paper's reminder as follows: Related works and comparisons between this paper and recent works are given in section 2, Section 3 provides theoretical background. Section 4 contains a discussion of the proposed methodology. After that, in section 5, there is a discussion of the findings. Finally, section 6 contains the final observations.

2. Related works

We bring some recent papers that offer several clustering approaches, which have been proposed to predict and analyze COVID-19 cases. And we compared these papers with the current paper.

(Poompaavai A., & Manimannan G., 2019)This paper attempts to identify the affected Union Territories and Indian states by COVID-19 using clustering of K-means method depend on the secondary sources of data collected from Indian Health and Family welfare Organization until March 24, 2020. It produced an effective result and visualized their work [8].

(Juniar Hutagalung, et al., 2020): This paper discusses the grouping of cases of COVID-19 and deaths using the K-Means clustering in Southeast Asia. And utilizing Rapid Miner tools, the data was clustered into various clusters using the K-Means Clustering Process. Data have been used are 2020 deaths from WHO for April of year 2020, statistics of country, and COVID-19 confirmed instances [9].

(Md. Zubair, et al. 2020): The authors have presented in this paper algorithm of an efficient clustering of K-means, which during COVID-19 has produced the clusters of various countries in a similar way, its model reduces the times of performance. Although their suggested model outperforms real-world circumstances, it may be slightly divergent if the dataset has a large or considerable number of cases [10].

(Sukma Sindi, et al., 2020): To show the COVID-19 patterns of spatial distribution, this paper utilizes the K-Medoids. They explained that the PAM is an algorithm of categorizing analytical sections. To obtain a group of k-clusters among the data that most demand an object in the data combination. The findings of their paper showed the new research of COVID-19 gathering created in Indonesia from different regions [11].

(Sanjay Kumar Sonbhadra, et al., 2020): This paper have been suggested parallel one-class support vector machines based on a new bottom-up target guided method for mining of COVID-19 articles, generated with the help of K-means clustering, DBSCAN, and HAC, which have been trained on the clusters of related papers, It had been offered that the suggested technique generated essential findings [12].

(Fitria Virgantari and Yasmin Erika Faridhan, 2020): This paper had been intended to map COVID-19 cases in Indonesia's provinces using K-means clustering method. It had been an initial attempt to inform the public and raise the awareness of the disease spread. It had been hoped that this study may assist towards optimal handling of the pandemic in Indonesia [13].

(Achmad Solichin and Khansa Khairunnisa, 2020): The K-Means methods were employed in this paper for producing a data distribution prototype application for patient Covid-19. Depend on the findings of this paper, the DKI Jakarta government is expected to be able to make strategic decisions regarding the reduced spread of the Corona virus in DKI Jakarta [14].

(Shashank Reddy Vadyala, et al. 2020): The authors of this paper propose a neural network of K-Means-LSTM to deal with precision and variance when estimating the number of cases of COVID-19 that have been reported in the conventional SEIR models. Their results aided policymakers and healthcare providers in efficiently provide and prepare services to deal with the scenario in these states during the coming weeks and days containing beds, intensive care facilities, and nurses [15].

In our suggested model, we fetch an improvement to both PAM (K-Medoids) and K-Means. We propose to collect the advantages from these algorithms like as running scalability, simplicity, insensitivity and effectiveness to outliers and noise to carry out best than all the above approaches for prediction and analysis COVID-19 cases. In particular, we achieve K-MP, an algorithm depend on k-means and PAM. Table 1 offers

the difference between the recent related papers and this paper in the terms of method, objectives, programming environment, and performance measures. Table 1 turned out that our paper is the best in terms of method and in terms of performance measures.

Table1. Comparison between this paper and recent papers

References	Method	Objectives	Programming Environment	Performance Measures
Poompaavai A., &Manimannan G., 2019	K-means	Identify the affected areas by COVID-19 in India	Weka	Accuracy
Juniar Hutagalung, et al., 2020	K-means	discuss the grouping of Cases and Deaths of COVID-19 in Southeast Asia	Rapid Miner	NV
Md. Zubair, et al. 2020	Efficient K-means	Analyze COVID-19	Not Valid (NV)	Excution Time
Sukma Sindi, et al., 2020	K-medoids	COVID-19 spread analysis in Indonesia	Rapid Miner	NV
Sanjay Kumar Sonbhadra, et al., 2020	k-means with DBSCAN and HAC.	Exploration COVID-19	NV	Precision, Accuracy, Recall Specificity, & F1-Score
Fitria Virgantari & Yasmin Erika Faridhan, 2020	K-means	Map COVID-19 cases in Indonesia's provinces	Java	Varaince ratio
Achmad Solichin and Khansa Khairunnisa, 2020	K-means	making strategic decisions in reducing the spread of the Corona virus in DKI Jakarta.	MySQL & PHP	Sum of Square Error (SSE)
Shashank Reddy Vadyala, et al. 2020	K-Means-LSTM	Prediction of COVID-19 Cases	Python	Root Mean Sequare Error (RMSE)
This Paper	Hyprid of K-Means and PAM	Prediction of COVID-19 Cases in Iraq	Python	Accuracy, BA, Precision, Recall, TNR, F1 score, MAE,RMSE, Excution time

3. Theoretical background

3.1. K-Means clustering

Clustering of K-means is algorithm that divides the number of sates into k clusters. Each observation is associated with the cluster with the closest mean. By reducing the sum of square error between corresponding cluster centroid and data, the grouping is done. The classifying of data is the main objective of K-mean [16] [17]. Figure1 shows the steps of K-means, the closeness between these points $E_{(i,j)}$ Depend on the ED is offered in equation1 [18].

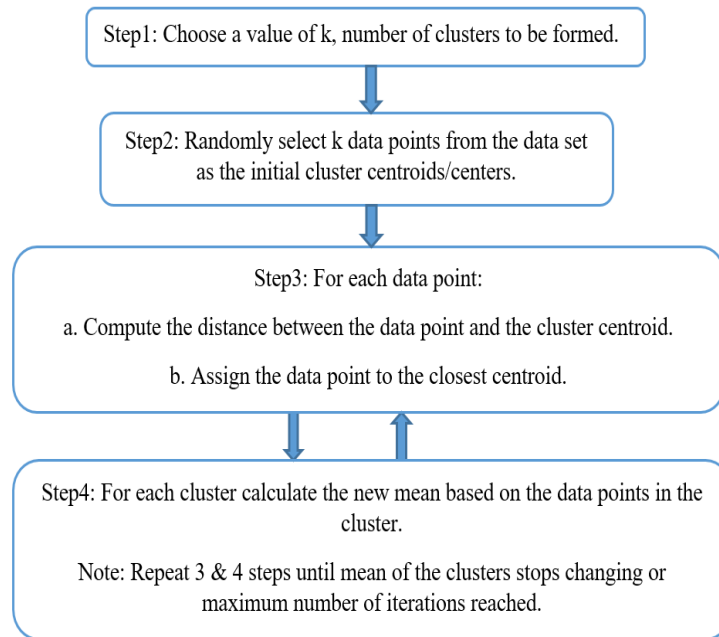$$E_{i,j} = \sqrt{(v_i - v_j)^T - (v_i - v_j)} \qquad (1)$$

Step1: Choose a value of k, number of clusters to be formed.

Step2: Randomly select k data points from the data set as the initial cluster centroids/centers.

Step3: For each data point:

a. Compute the distance between the data point and the cluster centroid.

b. Assign the data point to the closest centroid.

Step4: For each cluster calculate the new mean based on the data points in the cluster.

Note: Repeat 3 & 4 steps until mean of the clusters stops changing or maximum number of iterations reached.

Figure 1. Flowchart of K-Means steps

3.2. Partitioning around medoids clustering

PAM is a data division algorithm that separates s samples' data into k groups. In this algorithm, no relationship exists between (k+1) cluster solution and the k-group solution; consequently, it is a suitable grouping algorithm for big data. This algorithm is similar to clustering of k-means. Both aim to partition of data into k groups so that the groups reduce an error ESS criterion between a center and observation of the group. Equation 2 represents computation of ESS [19] [20]. Where c_k is the kth group (cluster) medoid and c(i) is the group consisting of xi. For more patterns, Figure2 describes the flowchart of the PAM [21] [22].

$$ESS = \sum_{r=1}^{k} \sum_{c(i)=k} \sqrt{(v_i - c_k)^T - (v_i - c_k)^T} \qquad (2)$$

Step1: Choose the proximity measure and the number of clusters k.

Step2: Form an initial assignment of the observations or points of the dataset used in the clustering into k clusters

Step3: Locate the medoid for each cluster.

Note: The medoid of the kth cluster is defined as that item in the kth cluster that

Minimizes the total distance to all other items within that cluster

Step4: For the rth cluster, where r=1, 2,…, k, reassign the ith

Observation to its nearest cluster medoid in such a way that

The objective function ESS presented in Eq. (2) decreases.

Note: Repeat the third and fourth steps until no further

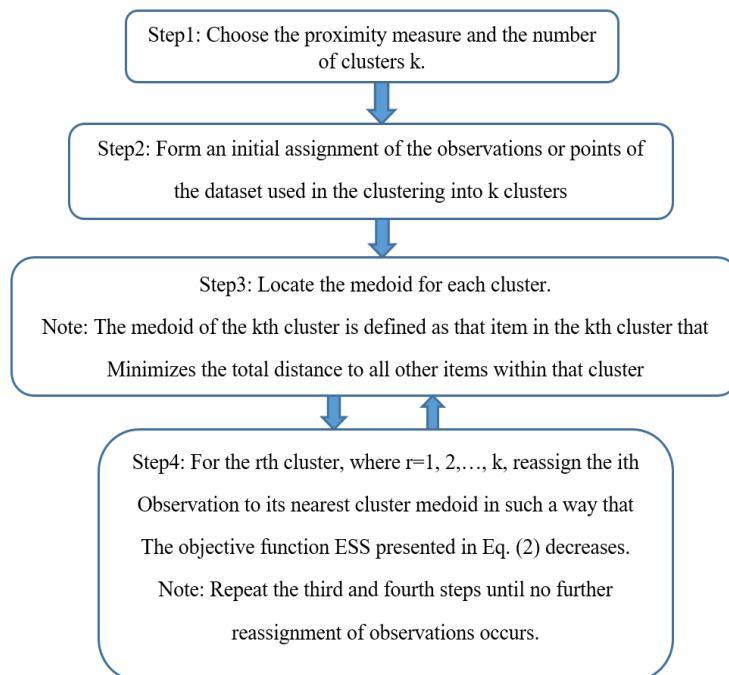reassignment of observations occurs.

Figure 2. Flowchart of PAM steps

3.3. A distinction between K-Means and PAM

A distinction between K-means and PAM is that the mean of observations is the center of k-means, commonly known as centroid as a reference point. On the other hand, the PAM selects a representative observation to serve as a reference point. To put it another method, the centroid in the k-means is the average or mean of the observations while the medoid of a cluster in the PAM is the most centrally located observation in the cluster that minimizes the total dissimilarity to all other observations during that cluster. As a result, PAM is more resistant to outliers and noise than k-means. . Figure 3 depicts the various between K-means and PAM [23] [24].
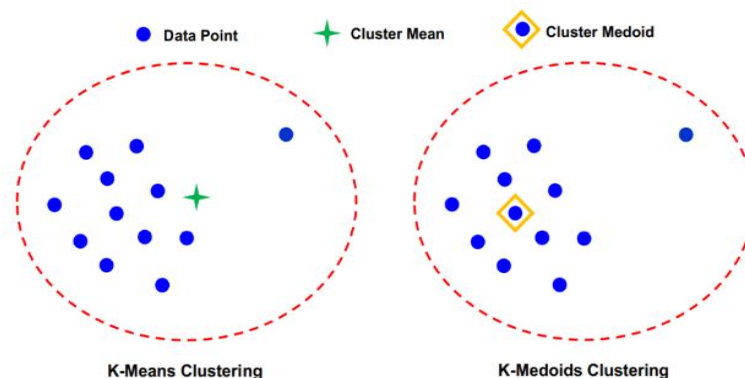
Figure3. Distinction between the PAM and K-Means

3.4. Performance measures

The prediction algorithms are evaluated by using typically measured using certain measures. In this paper, based on the confusion matrix the findings of forecasting are assessed by using measures such as precision, accuracy, precision, TNR, balance accuracy, f1 score, recall, and precision values. During the confusion matrix a complete view of the assessment of a predicting model is provided. It produces the findings of forecasting in the matrix form with the patterns of correctly forecasted cases, incorrectly forecasted cases, false of incorrect cases, and correct forecasting [25] [26]. We used the performance metrics of classification algorithms because we have the target class, so we treat it as the classification when evaluated it.

Accuracy: Accuracy of forecasting algorithm (ACC) is the rate of the actual class of dataset to the correct class of predictions. ACC of computation shown in equation2. Typically, any forecasting model produces four different findings, like as, False Negative (FN), True Positive (TP), False Positive (FP), and True Negative (TN). And the Balance Accuracy (BA) has calculated by applying equation3 [27] [28].

TPR is the Sensitivity, Recall or True Positive Rate: The number of correctly forecasted negative corona cases in the data set from all negative corona cases is the recall of the forecasting algorithm, the equation5 has applied for calculating recall. And the equation of Specificity, or TNR, Selectivity shown in (6). Precision: is the number of correctly predict corona negative cases that correspond to actual corona negative cases, as calculated by equation (6) [29].

F1 Score: It is both recall and precision measure of the balanced score (harmonic mean); equation7 used to calculate f1 score. Mean Absolute Error and Root Mean Square Error: Mean Absolute Error (MAE) is a measurement applied to calculate the mean of all the absolute value differences between the actual and the predicted cases. Computing the MAE is accomplished with equation9. RMSE calculates the average value of all the differences squared between the true and the predicted cases and then proceeds to compute the square root out of the finding. Computation of it given by equation 10. Where b_i : is the true cases. $\widehat{b_i}$ The predicted cases and n : are the number of cases [30] [31] [32].

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \tag{3}$$

$$BA = \frac{TPR+TNR}{2} \tag{4}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{5}$$

$$\text{TNR} = \frac{TN}{TN+FP} \tag{6}$$

$$\text{Precision} = \frac{TP}{TP+FP} \tag{7}$$

$$\text{F1 Score} = \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |b_i| - |\widehat{b_i}| \tag{9}$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (b_i - \widehat{b_i})^2} \tag{10}$$

4. Methodology

This work aims to improve the predictive model depend on similarity of patient by using K-Means, PAM, and K-MP to predict the negativity or positivity for COVID-19. Furthermore, we proposed using the K-MP technique to increase classification accuracy while reducing processing overhead. For implementing this aim, it is important to find a guide to adopt K-MP technique consists of major steps that consist of understanding and gathering of corona data, preprocessing and preparing of corona data, experiments and modeling, evaluating and testing. [33] [34] apply the idea of hybrid k-means and PAM but not on COVID-19 data.

The process in Fig.4 was stated as follows: We first collect and purify the data. We use the K-means, PAM, and K-MP algorithms, which include similarity calculations as a result of utilizing the K-means, PAM, and K-MP methods by using (1) and then the model predicts what the state is and continues to training those algorithms for every patient in a data collection; The status will be positive if the patient is infected with corona, and the patient's status will be negative if he or she is not infected with COVID-19. Finally, we assessed the suggested model by measuring Precision, BA, Recall, ACC, F1 Score & TNR depend on confusion matrix and MAE, RMSE & Execution time.
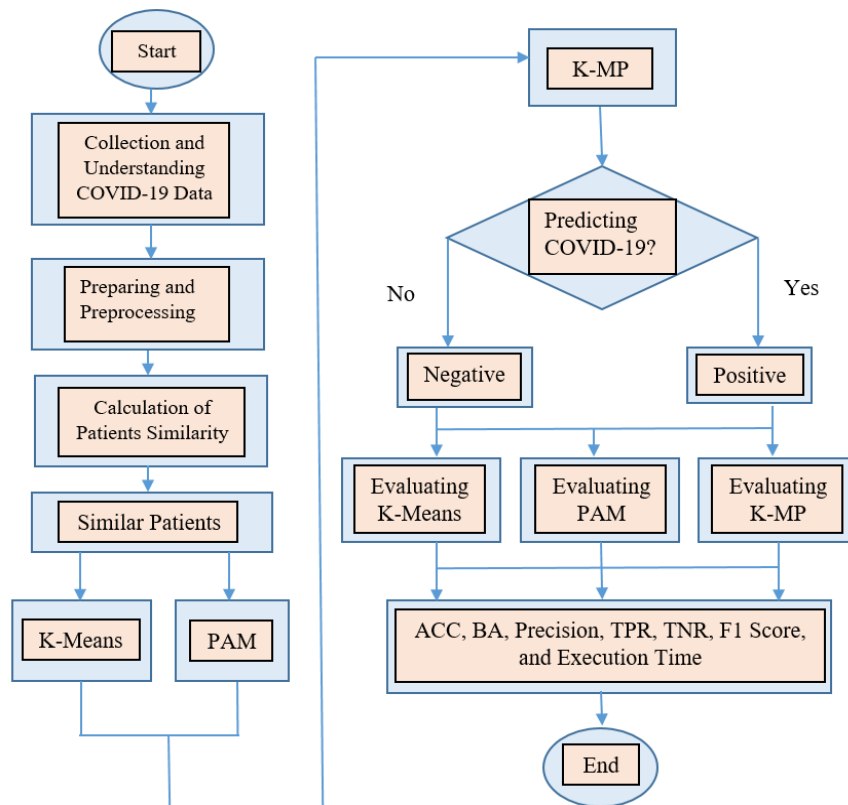

Figure 4. Flowchart of the suggested K-MP clustering for forecasting cases of corona

4.1. Gathering and understanding of corona data

In this research we will design a K-PM model to predict the status of corona for patients using actual data to obtain accurate and significant findings for decision-makers in the hospital. It must be a feasible method of gathering the relevant information. As a result, a questionnaire is created and distributed manually. There are several hospitals in Iraq that can affect and anticipate the status class (target). As shown in table2, the questioned features and symptoms for the training dataset are constrained between personal features and various laboratory test variables (symptoms) Questionnaire. These characteristics and symptoms are used to forecast whether the patient status is negative or positive. The questionnaire was completed by 400 patients from different hospitals in Iraq, with varying ages, genders, and governorates, in order to obtain complets patterns about them. This features was collected from the paper, which implementing by [35].

Table 2. The questionnaire about features and symptoms of patients

Question NO.	Description	Question NO.	Description
Q1	Age of the Patient	Q12	Is the Patient Suffer from Fatigue? (Yes, No)
Q2	Sex of the Patient	Q13	Is the Patient Suffer from Sputum? (Yes, No)
Q3	Governorate of the Patient	Q14	Is the Patient Suffer from Chest Pain? (Yes, No)
Q4	Nationality of the Patient	Q15	Is the Patient Suffer from Diarrhea? (Yes, No)
Q5	Temperature of the Patient (High, Medium, Low)	Q16	Is the Patient Suffer from Hypertension? (Yes, No)
Q6	Is the Patient Suffer from Cough? (Yes, No)	Q17	Is the Patient Suffer from Expectoration? (Yes, No)
Q7	Is the Patient Suffer from Headache? (Yes, No)	Q18	Is the Patient Suffer from Myaligia? (Yes, No)
Q8	Is the Patient Suffer from Loss of Smell Sense? (Yes, No)	Q19	Is the Patient Suffer from Gasp? (Yes, No)
Q9	Is the Patient Suffer from Loss of Taste Sense? (Yes, No)	Q20	Is the Patient Suffer from Somnolence? (Yes, No)
Q10	Is the Patient Suffer from Chronic Diseases? (Yes, No)	Q21	Is the Patient Suffer from Anorixia? (Yes, No)
Q11	Is the Patient Suffer from Dizziness ? (Yes, No)	Q22	What is the Status of COVID-19 for the Patient (Positive, Nogeative) This is the Target Class

4.2. Preparation and pre-processing of COVID-19 data

The procedure of feedback interpretation and analysis has been completed once the questionnaires were sent out and all the results collated, the raw corona data comprising non-applicable occurrences is processed as part of the COVID-19 data preparation procedure. The data set was cleaned using the pre-cleaning and processing techniques of data. This was due to inconsistencies and errors that needed to be corrected. COVID-19 data was imported to Excel sheets for evaluation and modification. The missing and outlier values are removed from the dataset throughout the data pre-processing and cleaning procedure. Therefore, data generalization is also seen as one of the approaches for reducing data. The Python was used to implement the model after the excel sheet was prepared and processed.

4.3. Algorithm of K-MP for prediction COVID-19 cases

For cases of COVID-19 prediction, we suggest K-MP, which is a technique that depends on the hybridization of PAM & k-means. Let us recall that k-means centroids might be fictive or actual objects, which are calculated as the cluster members' means. The algorithm of K-MP is summarized as follows:

Input: Count of COVID-19 clusters (K);

O a group of D objects;

Output: a group of K COVID-19 clusters;

Start

1: Call K-Means;

2: Compute the distance that separates each centroid (C) from the other d of the cluster;

3: Replace each C by the D that has the minimal distance;

4: Call PAM on the new C;

5: Replace the C by medoids computed from a randomly chosen subset of D, if better;

6: Continue to converge with k-means;

End;

5. Results and discussion

In this paper, applying the k-means and PAM algorithms of clustering to group patients, then the COVID-19 data is classified. Data were taken from the previous Iraqi hospitals used in this paper as a test case. First, the dataset is tested by ignoring the original label, then labeling the new data by grouping the data using the algorithm of k-means, using the PAM algorithm, and finally using the proposed K-MP algorithm.

By applying clustering techniques to the original dataset, divide the data into a suitable number of clusters, such as divide the data of 400 patients into 2 clusters according to the number of labels (positive, negative).

Next, predict the results of the status patient. Compare the products of K-Means, PAM, and K-PM of prediction. The result of dataset patient status, which was calculated using python programming language, can be seen in the below figures. The results are shown in terms of ACC, BA, recall, precision, TNR, and F1 score, which are calculated using formulas mentioned in the previous section of this paper. Also, we mentioned the results of execution time. For the evaluation and validation of three algorithms the sampling approaches are utilized. This approach analyzes how new undiscovered data on the algorithms may be predicted. The holdout is one of the methods used in this paper to validate the proposed model on the small data sample. As a result, clustering algorithms in this study take into account 30% of the test data and 70% of the training data. Figure5 refers to the testing findings for the ACC and BA for the three algorithms. We see that K-MP has the larger ACC and larger BA.

Figure6 shows the findings of the precision and recall measures for the three techniques on the COVID-19 data. We offer that K-MP has the best precision and recall. Figure7 illustrates the TNR and F1 score for the three algorithms. We remark that, the TNR and F1 score for K-MP are best from the K-Means and PAM. The experimental analysis Figure8 shows that the MAE and RMSE of the K-means will be higher than PAM and K-MP clustering. Figure9 offers the running time for the three techniques on COVID-19 data. We note that the time of execution for K-MP intermediates between the lower k-means and the highest PAM. Thus, for the benchmark, K-MP is successful and efficient in obtaining the best of both techniques, k-means' fast time and PAM's precision. Depend on the results, whichare shown above, the proper result of PAM is best than K-Means and the findings of K-MP are best than PAM. This means that the combination of PAM and K-Means is best than the other algorithm without combination for forecasting COVID-19 cases.
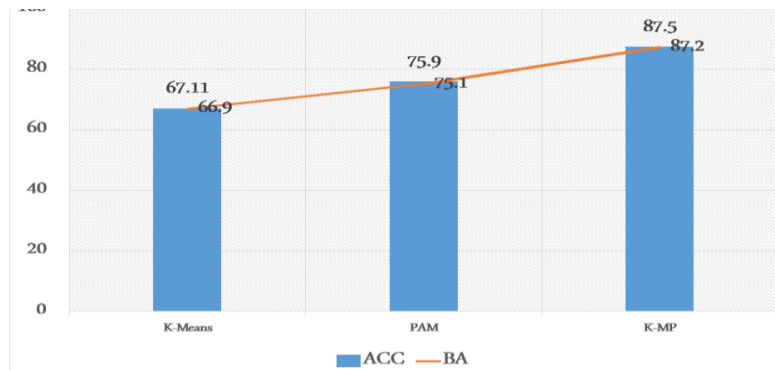
Figure 5. ACC and BA for K-Means, K-MP, and PAM
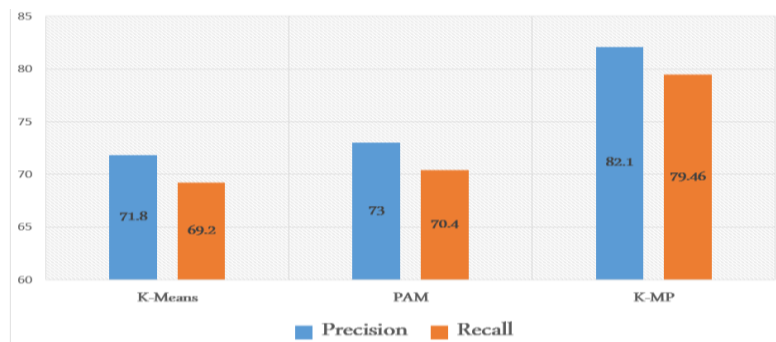

Figure 6. Precision and Recall for K-Means, K-MP, and PAM
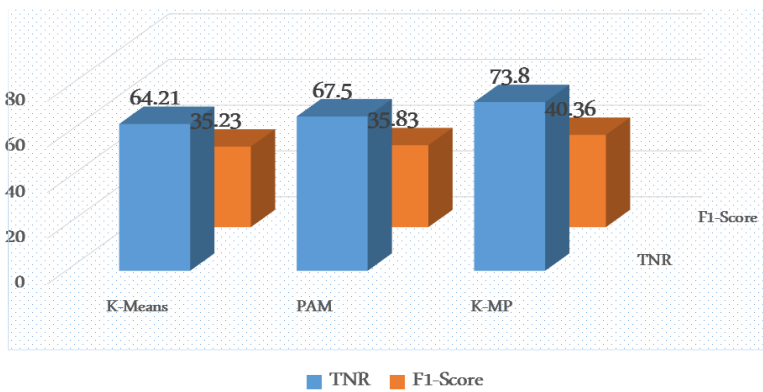

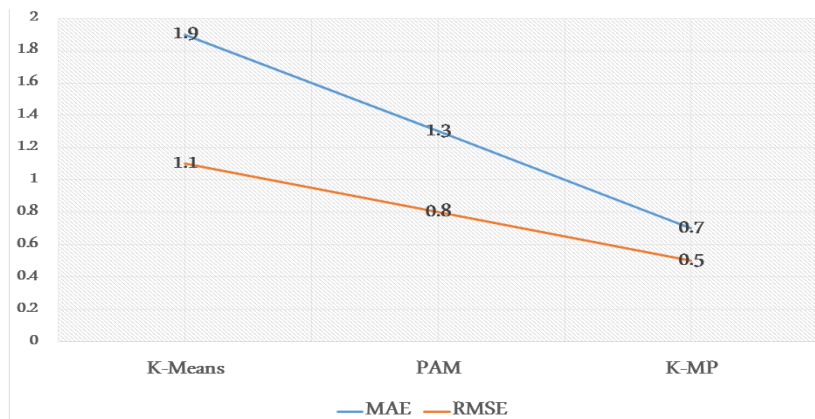Figure 7. TNR and F1 score for K-Means, K-MP, and PAM.

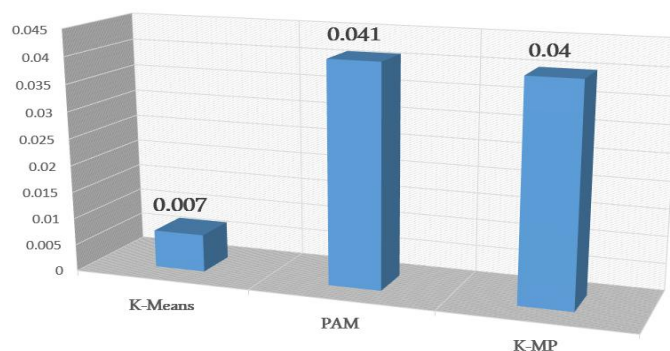Figure 8. MAE and RMSE for K-Means, K-MP, and PAM

Figure 9. Execution Time for K-Means, K-MP, and PAM

6. Conclusions

In this paper, we suggested integrating PAM and K-means clustering algorithms to increase effectiveness and efficiency in forecasting COVID-19 cases. The COVID-19 data set from Iraqi hospitals were used to test the experimental validity of the K-MP algorithm. K-MP results exceed k-means as well as PAM. It also outperforms recent papers in the literature. The findings of this paper may give feedback to hospitals in Iraq, allowing them to take even more concrete efforts to address the issue of Covid-19 transmission and death. Furthermore, the populace of each region in Iraq is becoming increasingly aware of Covid-19 transmission. Results show that the K-MP clustering algorithm achieved a satisfactory high ACC, Precision, Recall, TNR, MAE, RMSE, and F1 score compare to the k-means clustering and PAM clustering.

References

[1] F. Chen, et al., " Data Mining for the Internet of Things: Literature Review and Challenges," International Journal of Distributed Sensor Networks, Vol. 2015, Article ID 431047, 2015.

[2] S. Hussein Toman, et al., " Cluster-based information retrieval by using (K-means)- hierarchical parallel genetic algorithms approach," TELKOMNIKA Telecommunication, Computing, Electronics and Control, Vol. 19, No. 1, pp. 349~356. February 2021.

[3] A. Bazzi, B. Masini, A. Zanella, "Performance Analysis of V2V Beaconing Using LTE in Direct Mode with Full Duplex Radios," IEEE Wireless Communication Letters, vol. 4, no. 6, pp. 685-688, Dec 2015.

[4] Wang, et al., "Using Electronic Health Records and Machine Learning to Predict Postpartum Depression," Stud. Health Technol. Inform, Vol. 264, pp. 888-892, 2019.

[5] H. Drias, N. Fodil-Cherif and A. Kechid, k-mm: A hybrid clustering algorithm based on k-means and k-medoids, in: Advances in Nature and Biologically Inspired Computing, vol. 419 of the series Advances in Intelligent Systems and Computing, Springer International Publishing Switzerland, 2016, pp. 37–48.

[6] Fard, Maziar Moradi, Thibaut Thonet, and Eric Gaussier. "Deep k-means: Jointly clustering with k-means and learning representations." Pattern Recognition Letters 138 (2020): 185-192.

[7] Gupta, Manoj Kumar, and Pravin Chandra. "An empirical evaluation of K-means clustering algorithm using different distance/similarity metrics." Proceedings of ICETIT 2019. Springer, Cham, 2020. 884-892.

[8] A. Poompaavai and Manimannan Ganesan, "Clustering Study of Indian States and Union Territories Affected by Coronavirus (COVID-19) using k-means Algorithm , "International Journal of Data Mining And Emerging Technologies, 9(2), 43-51.

[9] J. Hutagalung, et al., "COVID-19 Cases and Deaths in Southeast Asia Clustering using K-Mean Algorithm. In Journal of Physics: Conference Series (Vol. 1783, No. 1, p. 012027). IOP Publishing, 2021.

[10] Md. Zubair, et al., "An Efficient K-means Clustering Algorithm for Analysing COVID-19". arXiv preprint arXiv:2101.03140, 2020.

[11] S. Sindi, et al., "Analisis algoritma k-medoids clustering dalam pengelompokan penyebaran covid-19 di indonesia" . JurTI (Jurnal Teknologi Informasi), 4(1), 166-173., 2020.

[12] S. K. Sonbhadra, "Target specific mining of COVID-19 scholarly articles using one-class approach". Chaos, Solitons & Fractals, 140, 110155. international journal of engineering and natural science, 2020.

[13] A. Solichin and K. Khairunnisa , "Klasterisasi Persebaran Virus Corona (Covid-19) Di DKI Jakarta Menggunakan Metode K-Means", Fountain of Informatics Journal, 5(2), 52-59, 2020.

[14] S. R. Vadyala, et al. ,"Prediction of the Number of COVID-19 Confirmed Cases Based on K-Means-LSTM" arXiv preprint arXiv:2006.14752.

[15] Fitria Virgantari and Yasmin Erika Faridhan, "K-Means Clustering of COVID-19 Cases in Indonesia's Provinces",

[16] Yu, Hao, et al. "Self-paced learning for k-means clustering algorithm." Pattern Recognition Letters 132 (2020): 69-75.

[17] N.C. Peeri, N. Shrestha, M.S. Rahman, R. Zaki, Z. Tan, S. Bibi, M. Baghbanzadeh, N. Aghamohammadi, W. Zhang, U. Haque, The SARS, MERS and novel coronavirus (COVID-19) epidemics, the newest and biggest global health threats: what lessons have we learned, Int. J. Epidemiol. 49 (3) (2020) 717–726, https://doi.org/ 10.1093/ije/dyaa033.

[18] M.A. Alzubaidi, M. Otoom, N. Otoum, Y. Etoom, R. Banihani, A Novel Computational Method for Assigning Weights of Importance to Symptoms of COVID-19 Patients, 2020. Under Review.

[19] G. Soni and A. Patel, "Comparative analysis of k-means and k-medoids algorithm on iris data," International Journal of Computational Intelligence Research, vol. 13, no. 5, pp. 899 – 906, 2017.

[20] M. Nikooghadam, et al. "COVID-19 Prediction Classifier Model Using Hybrid Algorithms in Data Mining." International Journal of Pediatrics 9.1 , 2021: 12723-12737.

[21] A. mit Kumar, et al. "A Novel K-Means Clustering and Weighted K-NN Regression Based Fast Transmission Line Protection." IEEE Transactions on Industrial Informatics , 2020.

[22] K. Mtonga, et al., "Machine Learning-Based Patient Load Prediction and IoT Integrated Intelligent Patient Transfer Systems," future internet, Vol. 11, pp. 236, 2021.

[23] P.Malarvizhi Kumar and U. Gandhi, "A novel three-tier Internet of Things architecture with machine learning algorithm for early detection of heart diseases," Computer& Electrong Engineering, Vol. 65, pp. 222–235, January 2018.

[24] F. Virgantari and Y. Erika Faridhan , "K-Means Clustering of COVID-19 Cases in Indonesia's Provinces", Proceedings of the International Conference on Global Optimization and Its Applications Jakarta, Indonesia, pp. 21-22, November 2020.

[25] Poompaavai A. and Manimannan G, "Clustering study of Indian states and union territories affected by coronavirus (COVID-19) using k-means algorithm", International Journal of Data Mining And Emerging Technologies, vol. 9(2), pp. 43-51, 2019.

[26] S. Jayita, C. Chandreyee and B. Suparna, "Review of machine learning and deep learning based recommender systems for health informatics", Deep Learning Techniques for Biomedical and Health Informatic, Studies in Big Data, vol. 68, pp. 101-126, 2020.

[27] S. Wang, et al., "Using Electronic Health Records and Machine Learning to Predict Postpartum Depression," Stud. Health Technol. Inform, Vol. 264, pp. 888-892, 2019.

[28] P. Govender and V. Sivakumar, "Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019), " Atmospheric Pollution Research, Vol. 11, pp. 40–56, 2020.

[29] A. Poompaavai and G. Manimannan," Clustering study of Indian states and union territories affected by coronavirus (COVID-19) using k-means algorithm, " International Journal of Data Mining And Emerging Technologies, Vol. 9, pp. 43-51, 2019.

[30] A. Sayed Abbas, et al.," K-Means and K-Medoids: Cluster Analysis on Birth Data Collected in City Muzaffarabad, Kashmir ," IEEE Access, Vol. 8, pp. 151847-151855, 2020.

[31] F. Rahman, et al. ," Application of Data Mining Technique using K-Medoids in the case of Export of Crude Petroleum Materials to the Destination Country ," IOP Conference Series: Materials Science and Engineering, Vol. 835, No. 1, p. 012058, 2020.

[32] C. Oktarina, et al., "Comparison of K-Means Clustering Method and K-Medoids on Twitter Data. Indonesian Journal of Statistics and Its Applications ," Indonesian Journal of Statistics and Its Applications, Vol. 4, pp. 189-202. 2020.

[33] H. Drias, A. Kechid, and N. Fodil-Cherif," k-MM: A Hybrid Clustering Algorithm based on k-means and k-medoids," In Advances in Nature and Biologically Inspired Computing (pp. 37-48). Springer, Cham, 2016.

[34] H. Drias, A. Kechid, and N. Fodil-Cherif, "A hybrid clustering algorithm and web information foraging," International Journal of Hybrid Intelligent Systems 13, 137–149 137, 2016, DOI 10.3233/HIS-160231.

[35] R. A. Jaleel, I. M. Burhan, and A. M. Jalookh , "A Proposed Model for Prediction of COVID-19 Depend on K-Nearest Neighbors Classifier: Iraq Case Study, " Proc. of the 3rd International Conference on Electrical, Communication and Computer Engineering (ICECCE) 12-13 June 2021, Kuala Lumpur, Malaysia, DOI: 10.1109/ICECCE52056.2021.9514171.