IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 25, 2020, accepted July 27, 2020, date of publication August 3, 2020, date of current version August 28, 2020.

*Digital Object Identifier 10.1109/ACCESS.2020.3014021*

# K-Means and K-Medoids: Cluster Analysis on Birth Data Collected in City Muzaffarabad, Kashmir

**SYED ALI ABBAS[1], ADIL ASLAM[1], AQEEL UR REHMAN[2], WAJID ARSHAD ABBASI[1], SAEED ARIF[3], AND SYED ZAKI HASSAN KAZMI[1]**

[1]Department of Computer Science and Information Technology, University of Azad Jammu and Kashmir, Muzaffarabad 13100, Pakistan
[2]Department of Electronics and Information Engineering, Southwest University, Chongqing 400715, China
[3]Department of Computer Science, Saudi Electronic University, Riyadh 11673, Saudi Arabia

Corresponding author: Syed Ali Abbas (ali.abbas@ajku.edu.pk)

**ABSTRACT** In the field of medical, each and every analysis is decisive as the study links to life of the subject under observation. One of the most vital area in the field of medical is the healthcare of expecting women in low income countries. High mortality rate due to increased number of caesarean section is evident because of poor medical infrastructure in the region, misunderstood religious teachings, low education and lack of proper decision making at the right time. The root cause analysis of situations demanding caesarean section is a tough job, however in the presence of historical data, one may extract useful information that will help supporting a medical decision by predicting the outcome. It is obvious that regional disparities have a huge impact on the residents of that region. A study performed on any region cannot be all applicable to the residents of some other distant region. This motive has established grounds to conduct a local study upon the data collected from expecting women in city Muzaffarabad, Kashmir. It is believed that the findings of this study will be significant for women that share more or less similar physical, social and maternal traits. Keeping this in mind, study presents an analysis of two clustering techniques for the investigation of appropriate algorithm that groups data into relevant clusters robustly. Firstly, we analyzed K-means and K-medoids algorithms' capability to cluster the data using different distance metrics. Secondly, data transformation techniques including scale, range and Yeo-Johnson are applied. Finally, transformed data are used in K-means and K-medoids algorithms' to generate cluster accuracy. It is observed that the results produced from transformed data are better than using raw data. Yeo-Johnson transformation method is found best for k-means (Hartigan & Wang), K-medoids (SEV distance function) and Rank k-medoids (SEV distance function) with mean accuracy 67.58%, 69.58% and 72.64% respectively.

**INDEX TERMS** Healthcare, machine learning, cluster analysis, K-medoids, k-means, caesarean section, birth data.

## I. INTRODUCTION

Healthcare is an attractive research domain of science because of its social implications. More advancement in healthcare consequently enhances the probability of healthy life. This idea has brought several interdisciplinary research outcomes that serve the said purpose in one way or another. Last decade has witnessed numerous prognoses and diagnoses based research articles targeting some disease or health issue. Most of the studies attempted to provide valuable details by performing classification or prediction of diseases that cause high mortality worldwide [1]. Several delicate domains, for example pregnancy complications, yet seek attention that is one of a major cause of death in low income countries as shown in Figure 1 [1]. Pregnancy complications cause higher number of deaths in lower middle income countries as compared to the number of deaths reported in advanced countries due to kidney diseases and breast cancer. Though, several attempts have been made to answer the complications during pregnancy and possible precautions, however, the generic findings are not entirely applicable to the people from different regions. Therefore, it is necessary

The associate editor coordinating the review of this manuscript and approving it for publication was Emre Koyuncu.

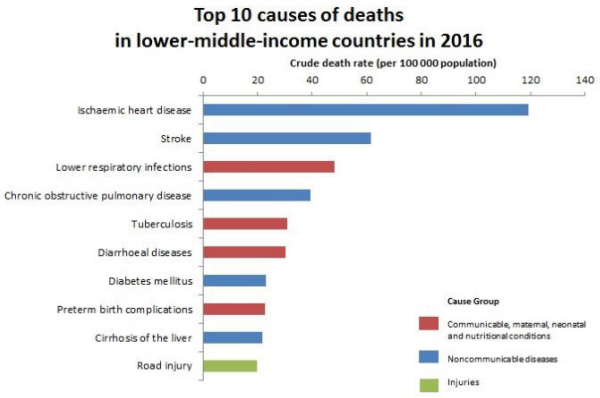



**FIGURE 1.** Top 10 causes of deaths in lower-middle-income countries.

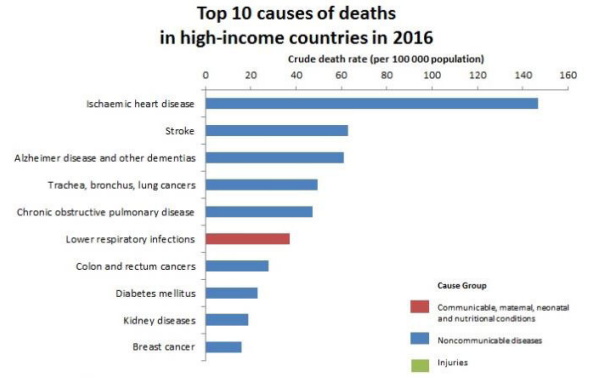



**FIGURE 2.** Top 10 causes of deaths in high-income countries.

to widen the sphere of research by conducting studies specific to the people sharing similar life styles. Consequently, the outcomes of such study will be significant and fully applicable to the people of that region. There are several factors that endanger expecting woman and her child. A lot of physical complications cause situations that lead to the need of Cesarean section (C-section), which itself is life threatening. The occasions that require C-section are birth of twins, triplets or more, a substantial infant, any previous birth by surgery, preterm births, diabetes etc. A C-section may be performed depending upon the shape of the mothers' uterus, history of a previous C-section, prolonged labor, cord prolapse etc. A C-section is often necessary when a vaginal delivery put mother or baby at risk [2]. More than 50 countries globally have C-section rates greater than 27 percent [3]. Efforts are being made to reduce occasions demanding C-section as the risk of death during surgical procedures is higher as compared to vaginal delivery. Therefore, the discovery of the factors that cause complications during pregnancy is essential. It is apparent that expecting mothers face more or less similar kinds of gestational, medical and physical experiences during pregnancy. These experiences are based on their social, physical and medical factors. The data of expecting women, if collected carefully, may become very handy for a physician to support his decisions based on the facts generated after sound data analysis. At this stage, data mining [4] and machine learning algorithms [5], [6] come into play which are excellent at learning patterns from data files and unveiling useful information that human eye fails to discover at once. It is believed that the development of decision support systems based on historical data, strengthened with learning capability of machine learning algorithms creates an opportunity for physicians to gain predictive information and take timely decisions that may save the life of expecting women and fetus. In computing, supervised and unsupervised learning algorithms are usually modeled to generate predictions through historic data. Clustering falls into unsupervised scheme of study that aims at grouping of instances into a cluster based on the instances perceived similarities [7], [8]. In current study, we present cluster analysis using k-means and k-medoids on birth data that has been collected from government hospitals of the city Muzaffarabad, capital of Azad Kashmir. Firstly, we analyzed K-means and K-medoids algorithms' capability to cluster the data using different distance metrics. Secondly, data transformation techniques including Scale, Range and Yeo-Johnson are applied to transform the data. Lastly, K-means and K-medoids algorithms' with different distance metrics are reused with transformed data to generate cluster accuracy. The application area of clustering is wide ranging covering major fields of sciences including image segmentation [9], [10], hand writing/object recognition [11], [12], big data mining [13], human genetic clustering [14], recommender systems [15], spatial data analysis [16], data reduction [17] etc. Clustering techniques are widely used in healthcare including cardio vascular diseases [18], [19], classification and prediction using cancer gene expression data [20], [21], diabetes [22], stroke [23], Alzheimer's [24] etc. In the light of literature, few contributions in maternal health/pregnancy complications domains are discussed below.

Median clustering method is used to establish a group with percentage data of maternal health services in Aceh Province, Indonesia [25]. Study describes the similarity in data, usage of distance metrics, process to enhance the distance metrics by the median clustering method, process to determine the number of clusters, interpret clustering results and to set grouping boundary to describe the characteristics of pregnant women's health services in each cluster [25]. Another cluster analysis on 33,740 preterm births data is performed to assess the determinants for various preterm birth sub types. Adapted k-means model and fuzzy algorithm were used to identify clusters using predefined conditions [26]. This study first considered K-means algorithm, followed by, the self-organizing map (SOM) technique to extract co-morbidity based clusters from a healthcare discharge data set. After validation of general cluster composition for diabetes mellitus, co-morbidity based clusters were identified for pregnancy. The SOM technique was found to infer distinct clustering of pregnancy ranging from normal birth to preterm birth, and potentially interesting comorbidities that could be validated by published literature. The promising results suggested that SOM technique is a valuable unsupervised clustering method for discovering co-morbidity based clusters [27]. In another

study, latent class cluster analysis is used to identify maternal exposure clusters and association between risk factor and birth defects. clustering was performed using adjusted odds ratios [28]. Study identified three latent maternal exposure clusters: a high-risk, a moderate-risk and a low-risk cluster. The use of intelligent data mining and cluster analysis is performed on the data of 222 pregnant women [29]. Cluster analysis gained 94.10% classification accuracy. Study claimed proposed method appropriate to classify expecting women during early trimesters. Taylor et. al., [30] studies the differences between the US women receiving prenatal care and others who does not. Discriminant analysis was used to evaluate the results gained after grouping women in clusters based on their similar characteristics. Study reported six replicable clusters of women which did not receive prenatal care [30]. Another study based on multiple imputation fuzzy clustering identified three exposure groups, non-exposed, lighter tobacco exposed, and heavier-tobacco exposed based on variables related smoking. Authors claimed that multiple imputation fuzzy clustering is good at categorizing patterns of exposure and their influence on results [31]. Few other interesting and relevant researches include preeclampsia disorder [32], [33], intrauterine disorders [34]–[36], Regional Disparities in Maternal and Child Health Indicators [37], effect of social and demographic factors on maternal health [38], [39], effect of facility birth maternal and perinatal mortality [40].
Literature witnesses the application of unsupervised methods for analyzing different dimensions of pregnancy and complications among expecting women. However, most of the studies are conducted in advanced countries of the world and their results are not entirely applicable upon women of different regions because of regional disparities [37]. Therefore, this research is carried out in Muzaffarabad, the capital city of Azad Jammu & Kashmir having population above one million. The city lacks in medical infrastructure, interdisciplinary liaisons and financial assistance to mitigate the pregnancy-based complications at once. However, it is believed that the success and continuity of such studies will help flourish grounds to improve healthcare infrastructure and uplift the society.

## II. METHODS

Current study aims to answer the issues that are associated to complications caused by C-section. Furthermore, the analysis and investigation of clustering methods and their robustness when utilised with our data set.To achieve this, we aligned our objectives as:

1) Collection of local data that reflects the traits of women sharing same social, economical and physical attributes.
2) Application of K-means and K-medoids clustering algorithms with different distance functions on collected data to analyze the clustering ability of said methods.
3) Analysis of behavior of K-means and K-medoids algorithms'with different distance metrics when provided with transformed data to generate cluster accuracy.

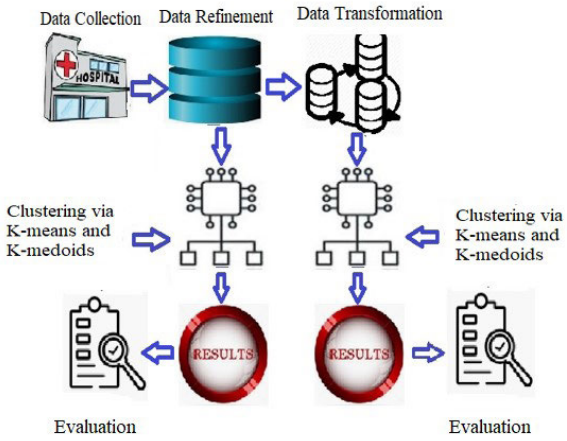



**FIGURE 3.** The scheme of the study.

The scheme of the study is depicted in Figure 3.

### A. ABOUT DATA

The data used is collected from the government hospitals of the city Muzaffarabad, the capital of Azad Kashmir. The Questionnaires were designed for data collection. While keeping in mind that most respondents come from rural areas with no educational background, questionnaires were filled in by a university graduate in the presence of obstetrician-gynecologist. Original data was comprised of 983 instances with 79 attributes divided into pregnancy, maternal history, medical condition, gestational and social life factors. The mode of the delivery (Caesarean or normal) is the outcome variable of the data. This variable makes the data suitable to conduct classification studies as well [41], [42]. Few attributes from the data set used in experiments are provided in table 1.

**TABLE 1.** Data Attributes of the Study.

| History | Medical/Physical | Social/Current Pregnancy |
|---|---|---|
| First Pregnancy | Increased breathing | Cousin marriage |
| Total pregnancies(Successful deliveries) | Increased heart beat | Education |
| Girls delivered | Headache | Working lady/Housewife |
| Boys delivered | Maximum blood pressure | Age |
| Aborption | Min blood pressure | Caesarean |
| Miscarriages | Haemoglobin | Normal |
| Last mode of delivery(caesarean/Normal) | Diabetic | – |
| Surgery(other than c section) | menstrual cycles (Regular/irregular) | – |

The attributes are in correlation with each another and effect the mode of the birth i.e., caesarean or normal. The correlation matrix of main attributes is provided in Figure 4. The key findings from data and correlation matrix are provided below.

1) Out of 983 cases, 571 cases belong to C-section and 412 cases to normal delivery.
2) Frequency of women with first pregnancy falls within age of 20 years to 27 years (312 out of 488).
3) Frequency of abortions is found to be highest among women between 18 years to 25 years age group (73 out of 112).

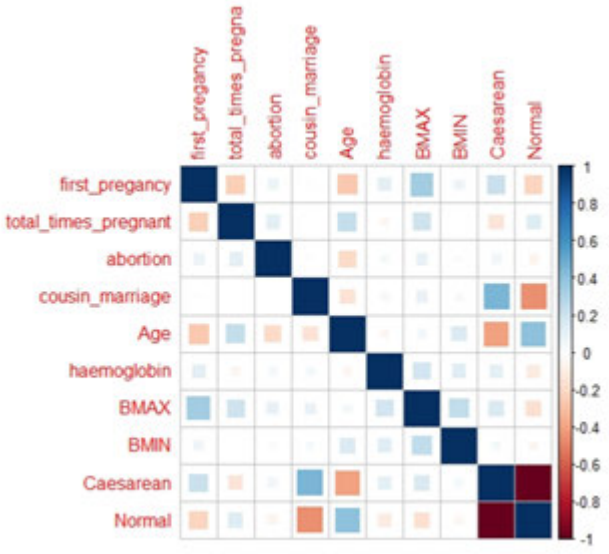



**FIGURE 4.** Correlation Matrix of main attributes used in study.

4) Women who have gone through surgeries (other than C-section) tend to deliver via C-section (35 out of 49 cases).
5) All women with C-section deliveries (most recent) have delivered via C-section in their current pregnancy.
6) Women with high and low blood pressure, lower levels of hemoglobin, diabetes and hypertension tend to deliver via C-section.
7) C-section rate is high among women of early ages (17-23) and late 30's.( 291 out of 374)
8) Cousin marriages are in high positive correlation with C-section deliveries (523 out of 682).
9) Working women category has higher number of C-section deliveries (197 out of 280).
10) High blood pressure is reflected among women with first pregnancy
11) Women with low educational background or no education at all delivered 4 children on average.
12) The number of girls delivered is almost double to the number of boys delivered.

### B. CLUSTERING

Clustering attempts to gather similar data into only one group and that data is not allowed to appear in any other group [26], [43]. In the presence of a good clustering method one may identify anomalies in data, gain more knowledge about data, generate hypothesis, discover rations of similarities in instances, perform compression etc. [44]. Clustering is a tough job that greatly depends upon the shape of the data. Presence of noise in data affects the shape, density and size of the cluster [44]. Human eye is good at clustering objects within two dimensions but with the increase of dimensionality, we need clustering algorithms that should provide isolated and compact clusters. There are several clustering methods available to be used in different conditions. Clustering is normally categorized as hierarchical clustering technique or partitional clustering technique [45]. Hierarchical clusters iteratively divide the patterns using bottom-up (agglomerative hierarchy) or top-down (divisive hierarchy) approach. In hierarchical clustering methods, clusters form by repeatedly dividing the patterns by using top-down or bottom up approach [46]. Contrary to hierarchical clustering, Partitional clustering attempts to classify the observations of the data into the $k$ clusters based on some criterion function. The criterion function most commonly attempts to find minimum distance between points in the available cluster. K-means [47], K-medoids [43], CLARA [43] etc., belong to this category. Clara is designed to handle data with several thousand objects, hence not utilized in current experiments.To investigate the efficacy of clustering methods with similar but large data set is left for the future. Though, the data set used in current study is small to medium sized, however practical application developed using proposed methodology will reduce the computational time while generating results. Further discussion this point forward is focused on K-means and K-medoids.

### C. K-MEANS

K-means is a simple algorithm that is famous for its effectiveness in clustering [47]. K-means attempts to minimize the squared error difference between the mean of the cluster and the data points in that cluster. Suppose we have some n-dimensional data points that a user wants to group in $k$ number of clusters with $\mu_k$ as a mean of that cluster, then k-means is represented as [44]:

$$\left[\sum_{k=1}^{k}\sum_{x\epsilon c_k}||x_i - u_k||^2\right] \tag{1}$$

where $x_i$ is the set of data points with $i = 1, 2, 3, .., n$, to be grouped in a cluster from a set of clusters given as $c_k$ with $k = 1, 2, 3, \ldots, k$. In order to reduce the squared error, k-means allocate patterns to initially partitioned $k$ clusters [44]. In case of non-decisive clusters' membership, k-means continue to repeat following steps [48].

1) Assign each pattern to its nearest cluster and generate new partition
2) Compute new cluster mean

Another important parameter required by k-means is distant metric. Typically, k-means is used with Euclidean distance metric that computes the root of squared differences between coordinates of objects. Euclidean distance is computed as follows:

$$distance(x, y) = \sqrt{\sum_{i=0}^{n}(X_i - Y_i)} \tag{2}$$

Other than Euclidean distance, Manhattan distance and Minkowski Distance metrics are available with same notion. Manhattan distance computes the absolute differences

between the points of pair of objects as follows.

$$distance(x, y) = |X_i - Y_i| \tag{3}$$

Minkowski distance is the generalization of Euclidean and Manhattan distance. It can be used with ordinal and quantitative variables. Minkowsky distance is depicted in equation 4.

$$distance(x, y) = \left[\sum_{i=1}^{n} ||x_i - x_i||^{\frac{1}{p}}\right]^{\frac{1}{p}} \tag{4}$$

Other famous distances include Mahalanobis distance to detect hyper ellipsoidal clusters [49] and Itakura–Saito distance [50] used in speech processing for vector quantization. In current study, k-means is applied with three different algorithms: the MacQueen *et al.* [47],Hartigan and Wong [51] and the Lloyd [52] algorithms.

### D. MacQueen ALGORITHM

Macqueen is an iterative algorithm that starts up with choosing the number of clusters, select distance metric and gets initial centers on the basis of some method. Macqueen then follow an iterative approach to assign case to a cluster on following condition. If the case is nearest to a centroid of a subspace that it belongs to, then no change is made. If the case is nearest to another centroid, then the case is reassigned to the new nearest centroid and the centroids of two effected clusters are recalculated.

### E. HARTIGAN AND WONG ALGORITHM (H AND W)

This algorithms' objective is an attempt to obtain the local optimal sum of squared errors with-in cluster. It means that H and W may assign a case currently residing in a cluster with the nearest centroid to any other subspace, with a condition that it will minimize the with-in cluster sum of squared error as shown in equation 5 [53].

$$SSE2 = \frac{N_i \sum ||x_i - C_i||^2}{N_i - 1} < SSE1 = \frac{N_i \sum_j ||x_i - C_1||^2}{N_1 - 1} \tag{5}$$

In equation 5, for all i≠1, if sum of square of the current cluster (SSE2) is smaller than the sum of square of another cluster (SSE1), then the case is assigned to SSE1 (new cluster) else it would be assigned to the current cluster.

### F. LLOYD ALGORITHM

Lloyd algorithm attempts to find a set of cluster centers for the n-dimensional data points such that $(x_1, x_2, x_3 \ldots .x_n) \epsilon d^n$ that is a solution to the minimization problem [53]:

$$E = \sum_{i=1}^{k} \sum_{j=1}^{n} d(C_i, x_{ij}) \tag{6}$$

Lloyd starts up with choosing the number of clusters, select distance metric and gets initial centers on the basis of some method. The iterative part is carried out in following 3 step fashions.

1) Assignment of each case of data set to a cluster based on a distance metric
2) Update centroid based on the mean value of cases assigned to cluster in previous step.
3) Repeat step 1 and step 2 until centroids stop changing.

### G. K-MEDOIDS

K-medoids or partitioning around the medoids algorithm is a variation in k-means algorithm, where data points are selected as medoids rather selecting mean as a centroid in k-means. A medoid can be considered an object within a cluster which has minimum average dissimilarity to the other objects of that cluster. The k-medoid algorithm begins with computing $K$ medoids and assigning each object of the dataset to the nearest medoid using some distance metric. Afterwards, k-medoids calculates the swapping cost for swapping object $P_i$ and medoid $M_i$ as follows.

$$COST_{PM} = \sum_{M_i} \sum_{p_i \epsilon m_i} |P_i - M_i| \tag{7}$$

When this cost decreases to a set threshold then following steps are taken by the algorithm.For each medoid M, data point P such that $P \neq M$,

1) 1. Consider the swap of M and P, and compute the cost change
2) If the cost change is the current best, remember this M and P swap.
3) Perform the swap of M and P. If, it decreases the cost then repeat step 1 and 2. Else the algorithm terminates

### H. RANKED K-MEDOIDS

This algorithm computes the similarity between pairs of objects once. The cost to update medoids in each iteration is $O(k*m)$ for $K$ clusters and m number of objects [54]. Ranked k-medoids (rkmed) follows following steps for assigning object to a medoid based on similarity.

1) Using distance metric, compute similarities for pairs of objects.
2) By sorting the similarity values, calculate R- matrix. In sorted index matrix, store the indexes of similar objects from most similar to least similar object.
3) Randomly choose a value for K.
4) From the sorted index matrix, choose the group of most similar objects to each medoid.
5) For every object in the group identified in step 4, calculate the hostility value using following equation

$$hV_i = \sum_{x_i \epsilon N} \gamma_{ij} \tag{8}$$

Where $X_i$ is an object in the set objects $N$ and $r_{ij}$ is the rank matrix.

6) The new medoid is the object with highest hostility value.
7) Reposition one of the medoids placed in the same group.

8) Go to step 4, until maximum iterations criterion meets.
9) Assign object to the most similar medoid.

### I. DISTANCE METRICS

The experimental tasks of current study are carried out in R, which is considered to be a comprehensive statistical computing tool supported with R language. There are several distance computing options are available for k-medoids in R [55]. Considering the type of data bset used in the study, we have incorporated Manhattan weighted by range (mwr), squared Euclidean weighted by range (ser) and squared Euclidean weighted by variance (sev) distance functions.

### J. DATA TRANSFORMATION

Heuristically, data transformation contributes in generating good results from machine learning algorithms. Some times altering the structure or format of the raw data makes it useful. In current study data is transformed using Scale, Range and Yeo-Johnson. In scaling, the standard deviation is calculated for an attribute and later each value is divided by this standard deviation to acquire transformed scaled data. Range transform is a normalization method where data is scaled into some provided range, i.e. [0, 1]. Third incorporated transformation method, Yeo and Johnson [56] is selected from power transform family. Yeo-Johnson shifts the distribution of data rather changing the values of data like in scale or range. In the presence of skewed data, Yeo-Johnson works effectively by shifting the distribution, that consequently reduces the skewness in data and increase models' performance.

## III. RESULTS & DISCUSSION

The aim of the study is to present an analysis of two clustering techniques to investigate appropriate algorithm that groups locally collected data into relevant clusters robustly. To achieve said objective, firstly we used k-means with H & W, MacQueen and Lloyd algorithms. The predicted value that falls within a cluster is then compared with actual class label to calculate the accuracy of clusters output. The mean accuracy of k-means with three algorithms is presented in Table 2. In order to analyze the clustering ability of k-means with transformed data, we repeated the same experiment after transforming data using scale, range and Yeo-Johnson. The mean accuracy for k-means with three algorithms is recalculated and presented in Table 2.

**TABLE 2. Mean Accuracy percentage of K-means with H & W, Lloyd and MacQueen with and without Data Transformation methods.**

| Setting | H & W | Lloyd | MacQueen |
|---|---|---|---|
| Without Data transformation | 61.88 | 60.04 | 60.32 |
| With Scale | 62.90 | 63.31 | 65.78 |
| With Range | 62.54 | 61.06 | 65.36 |
| With Yeo Johnson | 67.58 | 67.12 | 66.05 |

The results show that K-means works well with H&W algorithm by correctly clustering 61.88% of birth data (figure 5). The mean clustering accuracy for Lloyd and MacQueen algorithm is 60.04% and 60.32% respectively. The performance

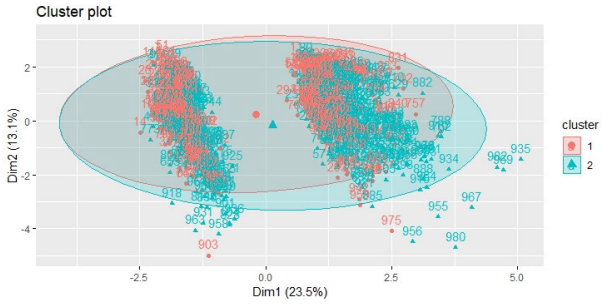



**FIGURE 5. K-means clusters actual vs. predicted without using transformed data.**

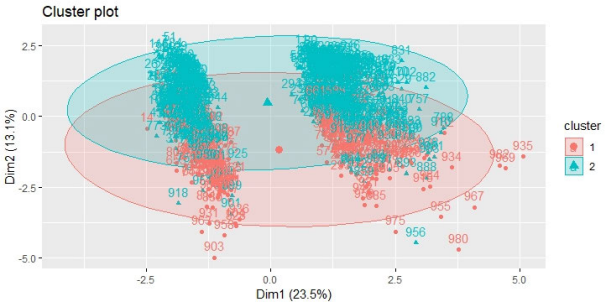



**FIGURE 6. K-means clusters actual vs. predicted with transformed data using YeoJohnson.**

**TABLE 3. Accuracies of K-medoids algorithm.**

| Setting | MWR | SER | SEV |
|---|---|---|---|
| Without Data transformation | 62.62 | 61.49 | 62.64 |
| With Scale | 64.54 | 63.57 | 65.59 |
| With Range | 64.59 | 64.59 | 65.01 |
| With Yeo Johnson | 67.67 | 68.03 | 69.58 |

of k-means clustering tends to improve with transformed data using scale, range and Yeo-Johnson. The highest percentage of correct clustering is achieved by H & W algorithm with Yeo-Johnson power transform i.e., 67.58% (figure 6). The mean accuracy of Lloyd and MacQueen algorithm with Yeo-Johnson transform is 67.12% and 66.05% respectively. The same procedure is repeated for K-medoids and ranked k-medoids using Manhattan weighted by range (mwr), squared Euclidean weighted by range (ser) and squared Euclidean weighted by variance (sev) distance functions. The mean accuracy of k-medoids and ranked k-medoids is presented in Table 3 and table 4 respectively.

K-medoids are better at clustering if compared with k-means, so far, this particular dataset is concerned. Without data transformation, k-medoids with squared Euclidean weighted by variance metric correctly clustered 62.64% of data. The mean accuracy of k-medoids using MWR and SER distance metrics is 62.62% and 61.49% respectively. Alike k-means, K-medoids clustering ability is improved with transformed data. Scale and range, though improves in terms of mean accuracy, however the difference in between scale and range is marginal. The highest percentage of correct clustering is achieved by k-medoids with SEV distance metric and Yeo-Johnson power transform i.e., 69.58%. Ranked

**TABLE 4. Accuracies of Ranked K-medoids algorithm.**

| Setting | MWR | SER | SEV |
|---|---|---|---|
| Without Data transformation | 61.42 | 66.02 | 65.69 |
| With Scale | 65.16 | 67.95 | 68.14 |
| With Range | 65.62 | 67.07 | 67.91 |
| With Yeo-Johnson | 68.87 | 69.27 | 72.64 |

k-medoids as the second variant of k-medoids is tested for its mean percentage accuracy of clustering. Results are provided in Table 4. The highest percentage of correct clustering is achieved using SEV distance metric and Yeo-Johnson power transform i.e., 72.64%.

The results clearly demonstrate the superiority of Yeo-Johnson transform when applied with ranked k-medoids incorporating squared Euclidean weighted by variance distance metric. This superiority is not accidental. The dataset has slight skewness that affects the outcome of all methods when used without transformation. Yeo-Johnson transforms the data by shifting its distribution, consequently helps in reducing the skewness and improving the results. The second goal i.e., to minimize the intra-cluster variance is achieved by SEV distance metric. The combination of Yeo-Johnson power transforms and squared Euclidean weighted by variance works well for the data set used in current study. The usefulness of these two techniques, when combined with k-medoids that minimizes the pairwise dissimilarities outperforms simple k-means.

## IV. CONCLUSION

It is observed that there are several studies conducted in advanced countries of the world. Their findings are significant, however, it is inferred that people as well the healthcare facilities differ from region to region, hence making the area specific research outcome inapplicable for people living in different parts of the world. Keeping this in mind, current cluster analysis based study using k-means and k-medoids clustering is carried out on locally collected birth data. Firstly; we analyzed k-means and k-medoids algorithms' capability to cluster the data using different distance metrics. Secondly, data transformation techniques including scale, range and Yeo-Johnson are applied and the clustering accuracy of the said clustering methods is re-calculated. It has been observed that the results after incorporating data transformation techniques are better as compared to the mean accuracy generated by clustering methods at alone. The highest percentage of correct clustering is achieved by rank k-medoids using SEV distance metric and Yeo-Johnson power transform i.e., 72.64%. It is believed that the predictions provided by any decision system based on reliable machine learning method will allow physicians to support their judgments and take effective measures to address some problem. Furthermore, several factors are associated that cause situations demanding C-section. For example, women married in early ages are at the highest risk to deliver via C-section. The trend of cousin marriages in the society should be addressed immediately as data reveals high number of C-section among women who married to their first cousins. High number of C-section among working women is alarming. The policy making departments should facilitate working women with stress free working environment, as stress itself is a huge contributor to other medical complications. Moreover, the number of newborn girls as compared to newborn boys is gradually increasing reflected by the data as well as in general. The policy makers should revise the rules for jobs and admissions in educational institutions that cater the needs of growing number of women. Contrary to a one published classification based study on same data set, presented study is first of its kind in a city having population above one million. It is believed, continuity of such interdisciplinary studies will help flourish grounds to improve healthcare infrastructure and uplift the society.

## ACKNOWLEDGMENT

The authors would like to thank the editor and anonymous reviewers for their perceptive comments and suggestions that helped us a lot to improve the quality of the article, which have been incorporated in this manuscript. They are thankful to Dr. Seemab Zafar (OB-GYN, Abbas Institute of Medical Sciences, Muzaffarabad) for her guidelines during data collection.

## REFERENCES

[1] World Health Organization. (2018). *The Top 10 Causes Death*. Accessed: Jun. 20, 2020). [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

[2] F. P. Guignard, "A gendered bun in the Oven. The gender-reveal party as a new ritualization during pregnancy," *Stud. Religion/Sci. Religieuses*, vol. 44, no. 4, pp. 479–500, Dec. 2015.

[3] G. Molina, T. G. Weiser, S. R. Lipsitz, M. M. Esquivel, T. Uribe-Leitz, T. Azad, N. Shah, K. Semrau, W. R. Berry, A. A. Gawande, and A. B. Haynes, "Relationship between cesarean delivery rate and maternal and neonatal mortality," *Jama*, vol. 314, no. 21, pp. 2263–2270, 2015.

[4] J. Han, J. Pei, and M. Kamber, *Data Mining: Concepts and Techniques*. Amsterdam, The Netherlands: Elsevier, 2011.

[5] A. L. Samuel, "Some studies in machine learning using the game of checkers," *IBM J. Res. Develop.*, vol. 3, no. 3, pp. 210–229, Jul. 1959.

[6] T. O. Ayodele, "Types of machine learning algorithms," *New Adv. Mach. Learn.*, vol. 3, pp. 19–48, 2010.

[7] A. Alsayat and H. El-Sayed, "Efficient genetic K-means clustering for health care knowledge discovery," in *Proc. IEEE 14th Int. Conf. Softw. Eng. Res., Manage. Appl. (SERA)*, Jun. 2016, pp. 45–52.

[8] A. K. Kar, S. K. Patel, and R. Yadav, "A comparative study & performance evaluation of different clustering techniques in data mining," in *Proc. ACEIT Conf.*, 2016, pp. 139–142.

[9] D. A. Forsyth and J. Ponce, *Comput. Vision: A Modern Approach*. Upper Saddle River, NJ, USA: Prentice-Hall, 2002.

[10] D. Lam and D. C. Wunsch, "Clustering," in *Academic Press Library in Signal Processing*, vol. 1. Amsterdam, The Netherlands: Elsevier, 2014, pp. 1115–1149.

[11] S. D. Connell and A. K. Jain, "Learning prototypes for online handwritten digits," in *Proc. 14th Int. Conf. Pattern Recognit.*, vol. 1, Jun. 1998, pp. 182–184.

[12] C. Dorai and A. K. Jain, "Shape spectra based view grouping for free-form objects," in *Proc. Int. Conf. Image Process.*, vol. 3, 1995, pp. 340–343.

[13] S. Arora and I. Chana, "A survey of clustering techniques for big data analysis," in *Proc. 5th Int. Conf.-Confluence Next Gener. Inf. Technol. Summit (Confluence)*, Sep. 2014, pp. 59–65.

[14] J. M. Kaplan and R. G. Winther, "Prisoners of abstraction? The theory and measure of genetic variation, and the very concept of 'race,'" *Biol. Theory*, vol. 7, no. 4, pp. 401–412, Jun. 2013.

[15] F. Ricci, L. Rokach, and B. Shapira, "Introduction to recommender systems handbook," in *Recommender Systems Handbook*. Springer, 2011, pp. 1–35.

[16] M. Honarkhah and J. Caers, "Stochastic simulation of patterns using distance-based pattern modeling," *Math. Geosci.*, vol. 42, no. 5, pp. 487–517, Jul. 2010.
[17] A. Hinneburg *et al.*, "An efficient approach to clustering in large multimedia databases with noise," in *Proc. 4th Int. Conf. Knowl. Discovery Data Mining*. Konstanz, Germany: Bibliothek der Univ. Konstanz, 1998, pp. 58–65.
[18] H. A. DeVon, C. J. Ryan, A. L. Ochs, and M. Shapiro, "Symptoms across the continuum of acute coronary syndromes: Differences between women and men," *Amer. J. Crit. Care*, vol. 17, no. 1, pp. 14–24, Jan. 2008.
[19] J. K. Herr, J. Salyer, M. Flattery, L. Goodloe, D. E. Lyon, C. S. Kabban, and D. G. Clement, "Heart failure symptom clusters and functional status–a cross-sectional study," *J. Adv. Nursing*, vol. 71, no. 6, pp. 1274–1287, Jun. 2015.
[20] H. Chipman, "Hybrid hierarchical clustering with applications to microarray data," *Biostatistics*, vol. 7, no. 2, pp. 286–301, Aug. 2005.
[21] M. C. de Souto, I. G. Costa, D. S. de Araujo, T. B. Ludermir, and A. Schliep, "Clustering cancer gene expression data: A comparative study," *BMC Bioinf.*, vol. 9, no. 1, p. 497, Dec. 2008.
[22] W. Chen, S. Chen, H. Zhang, and T. Wu, "A hybrid prediction model for type 2 diabetes using K-means and decision tree," in *Proc. 8th IEEE Int. Conf. Softw. Eng. Service Sci. (ICSESS)*, Nov. 2017, pp. 386–390.
[23] Y. Karaca, C. Cattani, M. Moonis, and Ş. Bayrak, "Stroke subtype clustering by multifractal Bayesian denoising with fuzzy c means and K-means algorithms," *Complexity*, vol. 2018, pp. 1–15, Apr. 2018.
[24] D. Gamberger, B. Ženko, A. Mitelpunkt, and N. Lavrač, "Homogeneous clusters of Alzheimer's disease patient population," *Biomed. Eng. OnLine*, vol. 15, no. S1, pp. 21–34, Jul. 2016.
[25] Z. M. Kesuma, Nurhasanah, and P. Kesuma, "Maternal health care in Aceh province: Cluster analysis results," *J. Phys., Conf. Ser.*, vol. 1116, Dec. 2018, Art. no. 022019.
[26] R. T. Souza *et al.*, "Cluster analysis identifying clinical phenotypes of preterm birth and related maternal and neonatal outcomes from the Brazilian multicentre study on preterm birth," *Int. J. Gynecol. Obstetrics*, vol. 146, no. 1, pp. 110–117, 2019.
[27] J. Chang and I. N. Sarkar, "Using unsupervised clustering to identify pregnancy co-morbidities," in *Proc. AMIA Summits Transl. Sci.*, 2019, p. 305.
[28] H. Cao, X. Wei, X. Guo, C. Song, Y. Luo, Y. Cui, X. Hu, and Y. Zhang, "Screening high-risk clusters for developing birth defects in mothers in Shanxi Province, China: Application of latent class cluster analysis," *BMC Pregnancy Childbirth*, vol. 15, no. 1, p. 343, Dec. 2015.
[29] I. Banjari, D. Kenjerić, K. Šolić, and M. L. Mandić, "Cluster analysis as a prediction tool for pregnancy outcomes," *Collegium Antropologicum*, vol. 39, no. 1, pp. 247–252, 2015.
[30] C. R. Taylor, G. R. Alexander, and J. T. Hepworth, "Clustering of us women receiving no prenatal care: Differences in pregnancy outcomes and implications for targeting interventions," *Maternal Child Health J.*, vol. 9, no. 2, pp. 125–133, 2005.
[31] H. Fang, C. Johnson, C. Stopp, and K. A. Espy, "A new look at quantifying tobacco exposure during pregnancy using fuzzy clustering," *Neurotoxicol. Teratol.*, vol. 33, no. 1, pp. 155–165, Jan. 2011.
[32] V. Sitras, R. H. Paulssen, H. Grønaas, J. Leirvik, T. A. Hanssen, Å. Vårtun, and G. Acharya, "Differential placental gene expression in severe preeclampsia," *Placenta*, vol. 30, no. 5, pp. 424–433, May 2009.
[33] K. Leavey, S. J. Benton, D. Grynspan, J. C. Kingdom, S. A. Bainbridge, and B. J. Cox, "Unsupervised placental gene expression profiling identifies clinically relevant subclasses of human preeclampsia," *Hypertension*, vol. 68, no. 1, pp. 137–147, Jul. 2016.
[34] S. D. Mahajan, R. Aalinkeel, S. Singh, P. Shah, N. Gupta, and N. Kochupillai, "Endocrine regulation in asymmetric intrauterine fetal growth retardation," *J. Maternal-Fetal Neonatal Med.*, vol. 19, no. 10, pp. 615–623, Jan. 2006.
[35] K. S. Bagi and K. S. Shreedhara, "Biometric measurement and classification of IUGR using neural networks," in *Proc. Int. Conf. Contemp. Comput. Informat. (IC3I)*, Nov. 2014, pp. 157–161.
[36] A. Zamecznik, K. Niewiadomska-Jarosik, J. Zamojska, J. Stańczyk, A. Wosiak, and J. Moll, "Intra-uterine growth restriction as a risk factor for hypertension in children six to 10 years old," *Cardiovascular J. Afr.*, vol. 25, no. 2, p. 73, 2014.
[37] E. Raheem, J. R. Khan, and M. S. Hossain, "Regional disparities in maternal and child health indicators: Cluster analysis of districts in Bangladesh," *PLoS ONE*, vol. 14, no. 2, Feb. 2019, Art. no. e0210697.
[38] I. Kalule-Sabiti, A. Y. Amoateng, and M. Ngake, "The effect of socio-demographic factors on the utilization of maternal health care services in Uganda," *Afr. Population Stud.*, vol. 28, no. 1, pp. 515–525, 2014.
[39] S. Pandey and S. Karki, "Socio-economic and demographic determinants of antenatal care services utilization in central nepal," *Int. J. MCH AIDS (IJMA)*, vol. 2, no. 2, p. 212, 2013.
[40] S. Gabrysch, R. C. Nesbitt, A. Schoeps, L. Hurt, S. Soremekun, K. Edmond, A. Manu, T. J. Lohela, S. Danso, K. Tomlin, B. Kirkwood, and O. M. R. Campbell, "Does facility birth reduce maternal and perinatal mortality in Brong Ahafo, Ghana? A secondary analysis using data on 119 244 pregnancies from two cluster-randomised controlled trials," *Lancet Global Health*, vol. 7, no. 8, pp. e1074–e1087, Aug. 2019.
[41] S. A. Abbas, R. Riaz, S. Z. H. Kazmi, S. S. Rizvi, and S. J. Kwon, "Cause analysis of caesarian sections and application of machine learning methods for classification of birth data," *IEEE Access*, vol. 6, pp. 67555–67561, 2018.
[42] S. A. Abbas, A. U. Rehman, F. Majeed, A. Majid, M. S. A. Malik, Z. H. Kazmi, and S. Zafar, "Performance analysis of classification algorithms on birth dataset," *IEEE Access*, vol. 8, pp. 102146–102154, 2020.
[43] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, vol. 344. Hoboken, NJ, USA: Wiley, 2009.
[44] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognit. Lett.*, vol. 31, no. 8, pp. 651–666, Jun. 2010.
[45] C. Fraley and A. E. Raftery, "How many clusters? Which clustering method? Answers via model-based cluster analysis," *Comput. J.*, vol. 41, no. 8, pp. 578–588, Aug. 1998.
[46] F. Murtagh, "A survey of recent advances in hierarchical clustering algorithms," *Comput. J.*, vol. 26, no. 4, pp. 354–359, Nov. 1983.
[47] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Statist. Probab.*, Oakland, CA, USA, 1967, vol. 1, no. 14, pp. 281–297.
[48] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*. Upper Saddle River, NJ, USA: Prentice-Hall, 1988.
[49] J. Mao and A. K. Jain, "A self-organizing network for hyperellipsoidal clustering (HEC)," *IEEE Trans. Neural Netw.*, vol. 7, no. 1, pp. 16–29, Jan. 1996.
[50] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. 28, no. 1, pp. 84–95, Jan. 1980.
[51] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *J. Roy. Stat. Soc., C Appl. Statist.*, vol. 28, no. 1, pp. 100–108, 1979.
[52] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. 28, no. 2, pp. 129–137, Mar. 1982.
[53] L. Morissette and S. Chartier, "The k-means clustering technique: General considerations and implementation in mathematica," *Tuts. Quant. Methods Psychol.*, vol. 9, no. 1, pp. 15–24, Feb. 2013.
[54] S. M. Razavi Zadegan, M. Mirzaie, and F. Sadoughi, "Ranked K-medoids: A fast and accurate rank-based partitioning algorithm for clustering large datasets," *Knowl.-Based Syst.*, vol. 39, pp. 133–143, Feb. 2013.
[55] W. Budiaji. (2019). *KMED: Distance-Based K-Medoids*. Accessed: Mar. 2020. [Online]. Available: https://cran.r-project.org/web/packages/kmed/vignettes/kmed.html
[56] I.-K. Yeo and R. A. Johnson, "A new family of power transformations to improve normality or symmetry," *Biometrika*, vol. 87, no. 4, pp. 954–959, Dec. 2000.

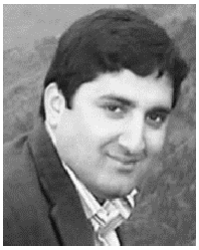
**SYED ALI ABBAS** received the Ph.D. degree in computer sciences from Chongqing University, China. He is currently serving as an Assistant Professor with the Department of CS&IT, University of Azad Jammu and Kashmir. His research interests include software metrics calibrations in software engineering, image encryption, and design and performance evaluation of machine learning algorithms.

**ADIL ASLAM** received the B.S. (CS) and M.Phil. degrees in computer sciences from the University of Azad Jammu and Kashmir, in 2012 and 2015, respectively. He is currently pursuing the Ph.D. degree with the Ankara Yıldırım Beyazıt University, Turkey. He is also a Research Associate with the Department of Computer Science and Information Technology, University of Azad Jammu and Kashmir. His research interests include machine learning-based decision support systems (DSSs), data mining, and artificial intelligence.

**AQEEL UR REHMAN** received the M.Sc. degree in computer science from the Islamia University of Bahawalpur, the second master's degree in computer engineering from UET Taxilla (CASE campus), Islamabad, Pakistan, and the Ph.D. degree in computer science and technology from Chongqing University, China. He worked as an Assistant Professor with the Department Computer Sciences, COMSATS Institute of Information Technology, Vehari Campus, Pakistan. He is currently working as a Senior Research Fellow with Southwest University, Chongqing, China. His research interests include non-linear dynamics and cryptography.

**WAJID ARSHAD ABBASI** received the Ph.D. degree in bioinformatics from the Pakistan Institute of Engineering and Applied Sciences (PIEAS), Islamabad, Pakistan on an HEC fellowship. He is currently a Faculty Member with the Department of Computer Sciences and Information Technology, University of Azad Jammu and Kashmir, Muzaffarabad, Pakistan. He has also been awarded the IRSIP by the Government of Pakistan for his research work with the Colorado State University, USA. His research interests include applications of machine learning in bioinformatics and the analysis of biomedical data.

**SAEED ARIF** received the M.S. degree in computer system engineering from the GIK Institute of Engineering Sciences and Technology and the Ph.D. degree from Hunan University China. He is currently an Assistant Professor with the Department of Computer Science, Saudi Electronic University Riyadh, Saudi Arabia. His research interests include digital forensics, digital watermarking, and machine learning.



**SYED ZAKI HASSAN KAZMI** received the Ph.D. degree in computer sciences from the University of Azad Jammu and Kashmir, Pakistan. He was a Programmer in software house. He was a Database Administrator with the Information Technology Board, Government of Azad Jammu and Kashmir. He is currently serving as an Assistant Professor with the Department of Computer Science and Information Technology, University of Azad Jammu and Kashmir. He has served public and private sectors in different capacities. His research interests include software engineering and biomedical signal processing.