**KING FAHD UNIVERSITY OF PETROLEUM AND MINERALS**

**College of Computer Sciences and Engineering**

*Computer Engineering Department*

COE 426 – Data Privacy Project (Term 201)

An Implementation of a Differentially Private K-Means Clustering Algorithm

|  |  |  |
| --- | --- | --- |
| **Student Name** | **Identification Number** | **Section Number** |
| Yousef Alfaris | 201443080 | 01 |
| Hussain Alsaffar | 201672140 |
| Abdullah Alnasser | 201535050 |
| Abdulaziz Alasiri | 201627560 |

|  |  |
| --- | --- |
| **Instructor** | Prof. Muhamed Felemban |

Table of Contents

[I. Introduction 2](#_Toc59741607)

[II. K-means & Differential Privacy 2](#_Toc59741608)

[III. Data set 4](#_Toc59741609)

[IV. Results 5](#_Toc59741610)

[V. Error Calculations 6](#_Toc59741611)

[VI. Conclusion 7](#_Toc59741612)

[VII.References 7](#_Toc59741613)

Table of Figures

[Figure 1 - Algorithm Variables ‎[3] 3](#_Toc59741600)

[Figure 2 – Pseudocode of the Algorithm ‎[3] 4](#_Toc59741601)

[Figure 3 - Elbow Method to Find Best K Value 5](#_Toc59741602)

[Figure 4 - 2D Graph of Basic K-Means Clustering Output 5](#_Toc59741603)

[Figure 5 - 2D Graph of DP K-Means Clustering Output using Epsilon = 70 6](#_Toc59741604)

[Figure 6 - Error Calculations over Different number of Epsilon 7](#_Toc59741605)

# Introduction

In this project, we will implement a K-means clustering algorithm that minimizes the risk of privacy disclosure. The proposed solution is to integrate differential privacy with K-means, which will help in reserving participants' privacy.

# K-means & Differential Privacy

K-mean algorithm is one of the simplest and popular unsupervised machine learning algorithms. It partitions a data set points into K distinct non-overlapping clusters where each point belongs to one group only ‎[1]. This can be used to confirm business assumptions about what types of groups exist or to identify unknown groups in complex datasets. Once the algorithm has been run and the groups are defined, any new data can be easily assigned to the correct group ‎[2]. However, K-means fail to consider the user’s privacy ‎[3].

Differential Privacy (DP) state that the analysis of any two neighbors’ datasets should be as close to each other. The neighbors’ dataset can be obtained by adding or removing a record of the original dataset. The similarity in the output of the analysis can be obtained by simply adding some noise with parameter ε , ε is the privacy budget in every query.

In this project, the idea of both K-mean clustering and DP are combined to interduce DP k-means clustering algorithm (DP-KCCM). Based on algorithm ‎[3], the dataset will be divided into n × k clusters, Initially the centroids C = n × k are selected using selection algorithm. Then, basic algorithm adds noise to each cluster and recalculates the centroids for a specific number of iterations. After the maximum iteration are reached, the algorithm merges the n × k centroids into k centroids. Figure 1 and Figure 2 show the variables used in the algorithm and the pseudocode of the algorithm, respectively.

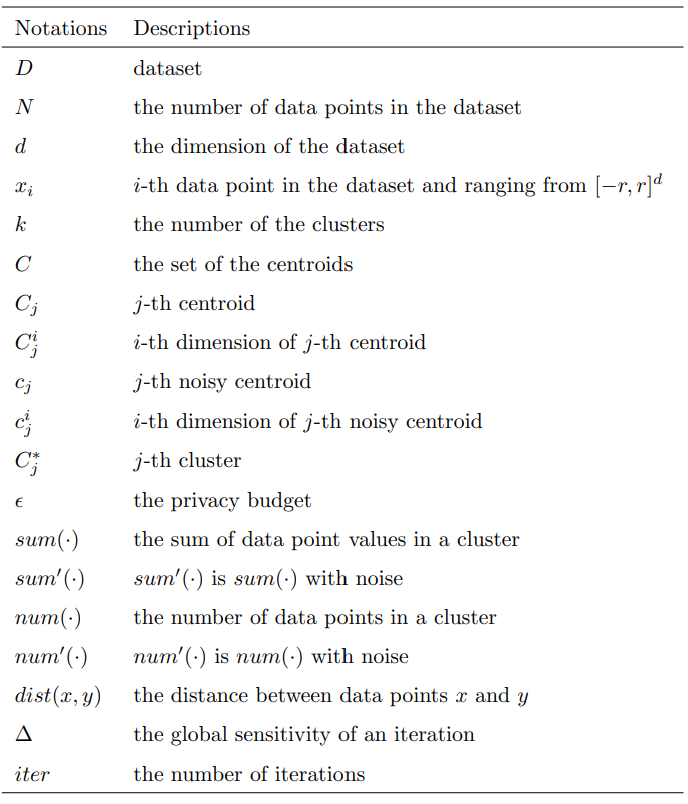


Figure 1 - Algorithm Variables ‎[3]

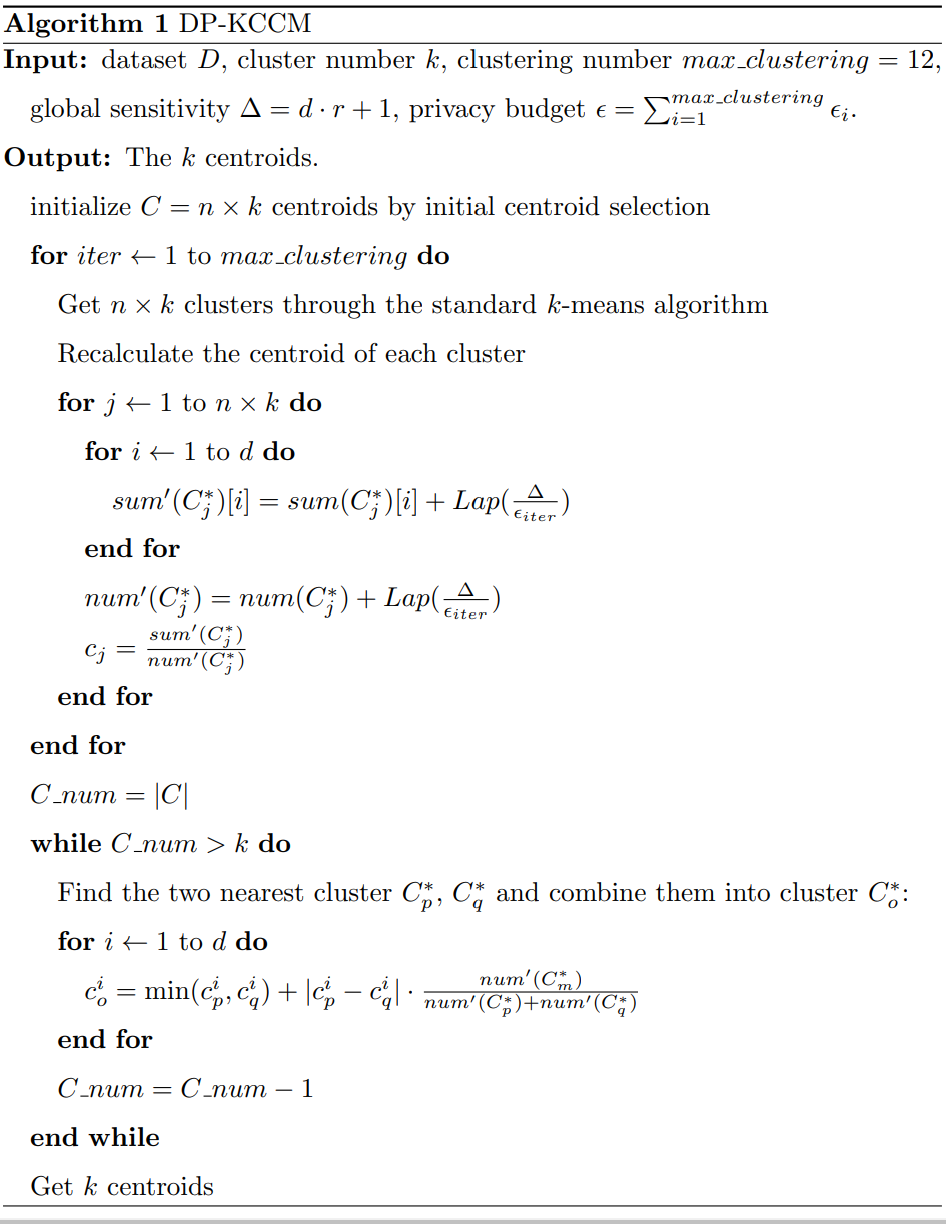


Figure 2 – Pseudocode of the Algorithm ‎[3]

# Data set

For the project, we have found four data sets that are suitable for unsupervised learning that contain somehow Personal Identifiable Information (PII). For the sake of simplicity and demonstration, we will apply the algorithm to dataset that shows information about customers of a mall. It has the attributes age, gender, annual income, and spending score. Currently, the data set has 200 records.

# Results

Choosing the optimal number of K is an important aspect in K-mean algorithm. So, we need an algorithm to find the optimal K based on the dataset. We choose to apply the elbow method which is one of the most popular methods to find the optimal value. as seen in Figure 3, a good value of K will be five (at the elbow), so we are going to have five clustering.

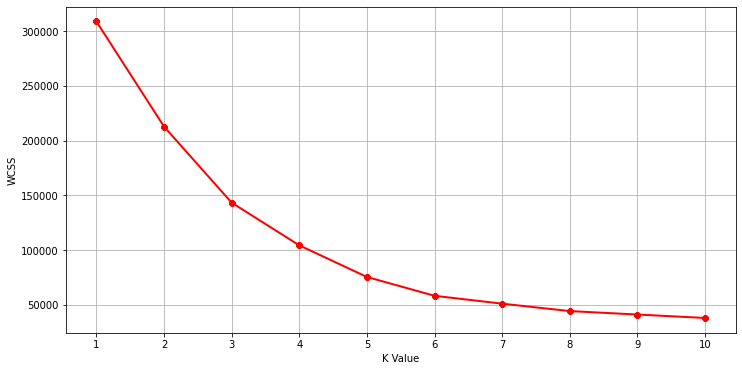


Figure 3 - Elbow Method to Find Best K Value

In Figure 4 we can see the output of original K-Means Clustering, since the optimal K was five and the centroids distributed between the data, we have five groups as shown below. Furthermore, the plots are done in a 2D form comparing each dimension in x-axis and y-axis. The output of the K-Means algorithm shows overlapping groups and how a sensitive attribute, Annual Income, is varying depending on other variables, such as Spending Score and Age.

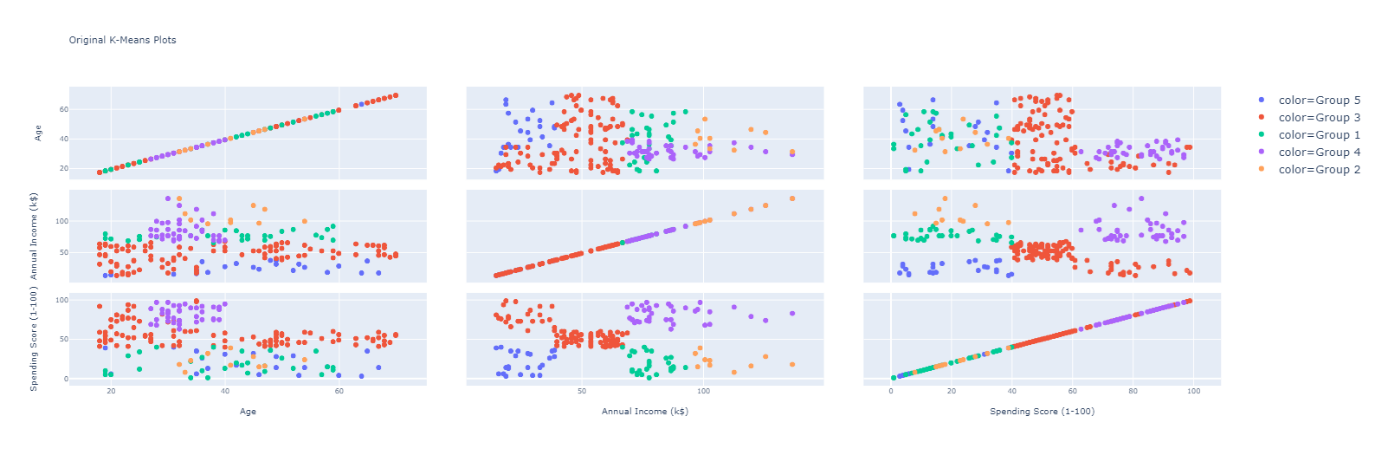


Figure 4 - 2D Graph of Basic K-Means Clustering Output

Figure 5 shows the output of dataset after DP K-Means using epsilon value of 70. We can see that it is very different from the original K-Means plot.

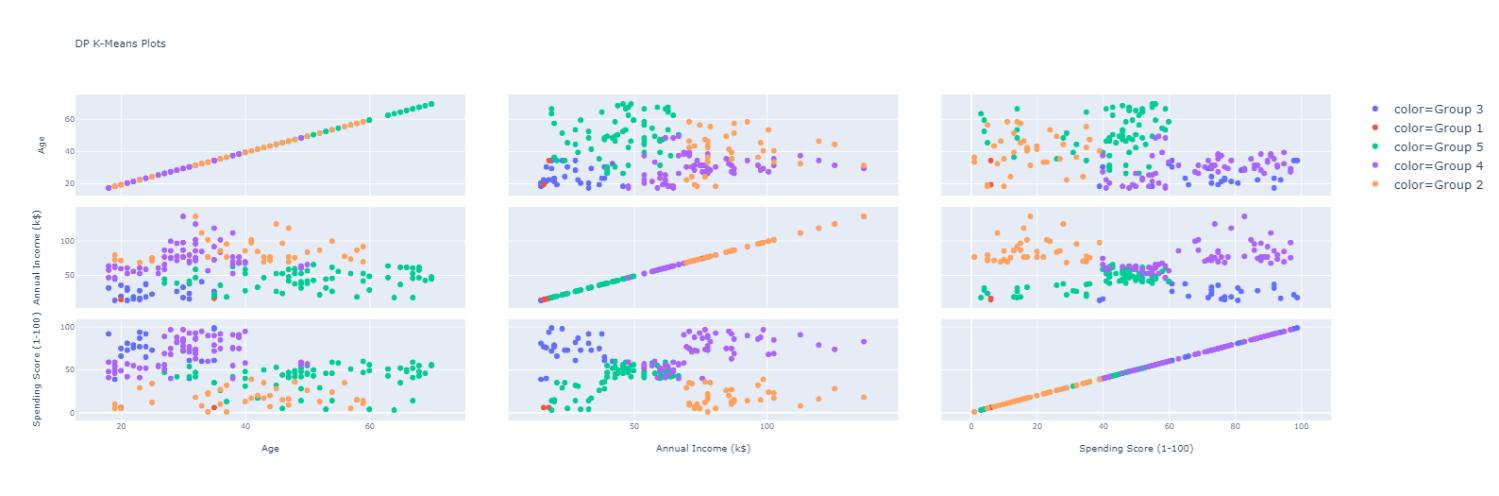


Figure 5 - 2D Graph of DP K-Means Clustering Output using Epsilon = 70

# Error Calculations

The dataset has been increased to 1500 records in order to get a more accurate estimation for the error. Based on our analysis, the error can be calculated using 3 approaches. First, it can be calculated by comparing the number of labels for the closest centroids between the original K-means and DP K-means. The idea to do this is simple, basically we compare the number of labels for each centroid, and then find the closest centroids. Based on that, we calculate sum of squared error. It will help in punishing the algorithm if there is a big difference in number of labels between the original and DP centroids. Downside of this algorithm is that it does not account for equal changing of corresponding records. The second method is calculating the error of distance difference between centroids. This can be done by finding the closest centroids, then finding the sum of squared error of their distance. Downside of this algorithm is the same as approach 1, it does not account for records of original dataset. Thirdly, a more complex and effective method can be used is to compare the number of changing records for a specific centroid. This needs the implementation of complex algorithm that checks the records in a group and compare them to closest group in the set of differentially private group. We have chosen the first and second approaches since they work greatly with our data set. In this section, we will show the result of first approach only. The function that generated the centroids grouped the dataset into 5 centroids, each one of these centroids have a set of labels. The algorithm mainly compares each label length of K-means and DP K-means, then, the centroids that have the closest lengths are used to get error value. The process is repeated until all centroids are removed and compared. Figure 6 shows the graph of Error vs Epsilon. To demonstrate the error correctness, we can that for small values of Epsilon, the error is very high, while the opposite for bigger numbers. The saturation Epsilon point of the error for our data set with the used settings in the algorithm is 25.

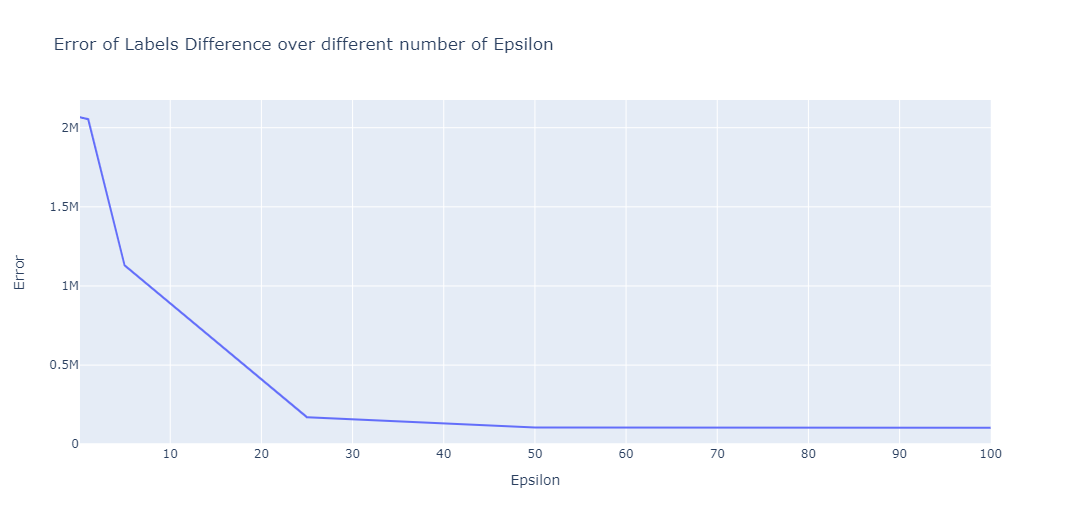


Figure 6 - Error Calculations over Different number of Epsilon

# Conclusion

Combining both concepts of K-means and DP can preserve the privacy of individuals while keep the benefits of getting statistical data. Applying this concept is neither hard nor impossible to achieve, so companies that have data about individuals should use similar approaches to ensure the privacy for their clients.

# References

1. I. Dabbura, “K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks,” *Medium*, 10-Aug-2020. [Online]. Available: https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a.
2. D. M. J. Garbade, “Understanding K-means Clustering in Machine Learning,” *Medium*, 12-Sep-2018. [Online]. Available: https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1.
3. T. Ni, M. Qiao, Z. Chen, S. Zhang, and H. Zhong, “Utility-efficient Differentially Private K-means Clustering based on Cluster Merging,” *arXiv.org*, 03-Oct-2020. [Online]. Available: https://arxiv.org/abs/2010.01234.