Clustering Analysis of Contraceptive Method Choices in Indonesia: An Advanced Machine Learning Approach

Joshua Bourton, *Member, IEEE*

Abstract—Globally, unintended pregnancies bring about considerable challenges for families and the communities they're a part of. The effects of unplanned parenthood reach far beyond the immediate challenges parents face, with women often finding themselves having to deal with heightened stress, heavy emotional struggles, and, in severe instances, post-natal depression as their lives take a sudden and dramatic turn. Studies conducted by Swarchz et al., (2008) have shown that inadequate family planning resources can push individuals to resort to unsafe abortion practices, posing grave dangers to maternal health.
These dangers are even more prevalent in regions where access to quality healthcare and comprehensive reproductive education is limited. Research by Gafar et al., (2020) has shown that in resource-constrained environments like these, a culmination of factors including limited healthcare infrastructure and societal stigmas surrounding reproductive choices create a perilous landscape for women. Thus, understanding the web of socioeconomic factors that contribute to unintended pregnancies becomes imperative for developing targeted interventions that can alleviate the disproportionate burden faced by individuals in such vulnerable settings.

Index Terms—Clustering Analysis, Contraceptive Method Choices, K-Means, Hierarchical

1 Introduction

This machine learning project trains two models using a subset of the 1987 National Indonesia Contraceptive Prevalence Survey donated by Lim (1997) to identify distinct sociodemographic groups with varying degrees of contraceptive non-use. The output for each model is a group of clusters representing distinct sociodemographic groups for the individuals in the survey.

Using the workflow shown in Figure 1, the model splits the dataset into distinct communities by analysing the following socio-demographic factors: Wife's Age, Number of Children (Numerical), Wife and Husband's Education, Husband's Occupation, Standard-of-Living Index (Categorical), Wife's Religion, Wife's employment and Media Exposure (Binary).

1.1 Significance

The distinct clusters generated through the K-Means and Hierarchical clustering methods (described below) successfully partitioned the dataset into distinct sociodemographic groups to uncover patterns within the dataset that most contribute to high unintended pregnancy rates. By identifying communities at higher risk of not using contraceptives, the model shows that high unintended pregnancy rates are in part due to inadequate access to and awareness of contraceptive methods in specific demographic groups.

Overall, K-Means clustering was the most efficient method, presenting more evenly distributed cluster means and higher Calinski-Harabasz index. Consequently, it stands out as the most suitable choice to be used in future supervised learning projects utilising the resulting clusters. The complex data-rich approach of hierarchical clustering, while ambitious, failed to surpass K-Means in performance. This may be attributable to the complicated interplay between dataset features and the nuanced social underpinnings influencing them, emphasising the challenges present in leveraging hierarchical clustering on real-world data.

However, when interpreting K-Means cluster significance, it's important to note the date at which the dataset was published (1987) and the limited data within it available for clustering, as the complexities involved in the features have no doubt changed over time since the release of the dataset.

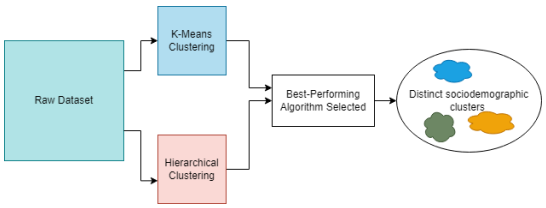


Fig. 1: Machine learning workflow for the project

2 Data and Experimental Setup

2.1 Dataset Characteristics

The distribution of the target variable (contraceptive method use) exhibited a notable imbalance in the dataset, with a predominant proportion of couples not using contraceptives, and about half as many individuals using short-term contraceptives. Upon examining the dataset through a heatmap (Shown in Figure 2), strong positive correlations were apparent between the husband's and wife's education levels, as well as the wife's age and number of children. These observed relationships offer valuable insight into potential sociodemographic features likely to emerge within distinct clusters through clustering analysis.

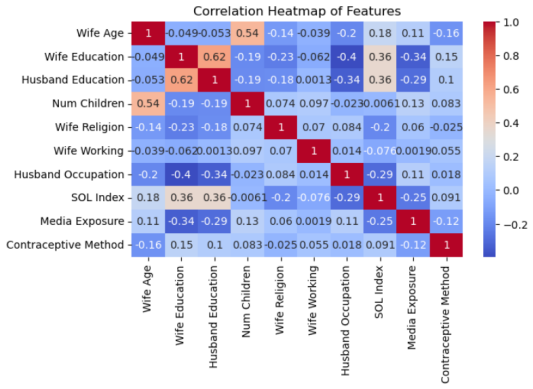


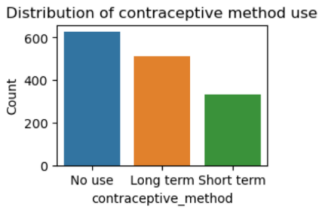
Fig. 2: Heatmap showing the correlation between features

Looking at the summary statistics for the numerical features, a significant variation in the number of children among couples becomes apparent, with a range of 0-16. This disparity underscores the necessity for a nuanced examination of clustering outcomes, with a particular emphasis on the variable that captures the number of children.

Additionally, the wife's age also reveals interesting insights. With a mean of 32 years, the women in the dataset exhibit a relatively youthful demographic profile, a factor which will exert a discernible impact on the clustering outcomes. The age distribution spans from 16-49 years, encapsulating a broad range of generational attitudes and life experiences. This diversity in age adds a nuanced layer to the analysis, offering a potential avenue for discovering unique traits within clustered communities.

	w_age	num_children
count	1472.000000	1472.000000
mean	32.544158	3.261549
std	8.227027	2.359341
min	16.000000	0.000000
25%	26.000000	1.000000
50%	32.000000	3.000000
75%	39.000000	4.250000
max	49.000000	16.000000

(a) Table summarising numerical features, showing wide variation in certain features



(b) Bar chart showing contraceptive method count

Fig. 3: Analytical figures showing disparities in the dataset

2.2 Data Preparation

To begin, the dataset was split into the training data and the target variable to exclude the target variable from clustering analysis. To allow for the plotting of multiple features on various graphs, the column names were first shortened. Z-score normalisation was then applied to the dataset to ensure that the clustering algorithm produced reliable and effective results. As discussed by Patel, Meta (2011), K-Means suffers when the features with the largest scales disproportionally influence the positions of the cluster centroids.

Normalising the data harmonised the fluctuations in scale between the various data types present in the dataset while mitigating the impact of outliers. Furthermore, the utilisation of encoded categorical variables was paramount for running the clustering algorithms, allowing for the comprehensive information encapsulated in categories to be represented whilst meeting the numerical requirements for calculating Euclidean distance in K-Means.

To visualise clusters effectively and facilitate human interpretation of the K-Means result, Principal Component Analysis (PCA) was applied. PCA serves as a dimensionality reduction technique, condensing data into just two dimensions and thereby enabling a more accessible representation of clustering outcomes. As discussed in research by Ding (2004), PCA serves to accentuate dominant patterns and structures within the clustered data, providing a concise and informative portrayal of the results.

2.3 Evaluation Metrics

For model comparison, various metrics were reviewed based on measuring cluster compactness and separation. The metric settled on was the Calinski-Harabasz index which, as described in research by Xu. W et al., (2019) measures the ratio of between-cluster to within-cluster variance, with a higher score indicating better-defined clusters. This provided a neat numerical way to evaluate the reliability of information extracted from the clusters.

Nevertheless, it's crucial to acknowledge the Calinski-Harabasz's sensitivity to cluster size. K-Means clustering tends to produce clusters of similar size, a trait hierarchical clustering often lacks. Consequently, the effectiveness by which the index measures the success of both methods may therefore not fully reflect the value that can be extracted from the resulting clusters.

Another insightful metric used was the mean of the target variable. Analysing the dispersion of these means helped to determine the extent to which distinct, usable clusters were formed. A wider spread of means implies variability amongst clusters, serving as an indicator of the model's success.

3 Model Evaluation

3.1 Choice of Machine Learning Algorithms

In the pursuit of differentiating between sociodemographic groups based on individuals' circumstances, two distinct clustering algorithms were selected. The first model employed, serving as the baseline, was the K-Means clustering algorithm. Renowned for its simplicity and efficiency, K-Means partitions the dataset into distinct groups, offering a foundational benchmark against which the performance of subsequent algorithms can be evaluated.

To complement and attempt to surpass the baseline, the second algorithm selected was the Hierarchical Clustering model. This choice stems from its ability to construct a hierarchical structure of clusters, presenting a more nuanced and layered perspective on the sociodemographic groups produced by the model. The hierarchical nature of this approach facilitates a more comprehensive exploration of how different socioeconomic attributes affect each other, thereby

enhancing the depth and usefulness of the derived clusters. The clarity of K-Means combined with the information-rich categorisation of Hierarchical Clustering yields a comprehensive understanding of the sociodemographic landscape under investigation.

3.2 Training Process

To uphold a robust experimental methodology, a particular emphasis was placed on the selection and fine-tuning of hyperparameters. One such hyperparameter is the number of clusters in K-Means. Both insufficient and excessive cluster numbers can lead to slow convergence rates thereby increasing computation time. If clusters are too few, results may not effectively capture the underlying patterns in data, causing underfitting as clusters emerge that are too broad and generalised to be interpreted effectively. Similarly, if the cluster number is too high, clusters will be too densely packed to depict meaningful distinctions among sociodemographic groups.

To find the optimum cluster number, traditional methods such as the generation of an elbow plot were initially employed. Candidate values were from one to ten clusters, as more than ten would take too great a time to interpret. However, this method yielded no discernible 'elbow point'. In light of this, an evaluation metric as discussed in research by Kodinariya et al., (2013) was employed - the silhouette score. As depicted in Figure 4, this plot demonstrated that the highest cluster number associated with the silhouette score was six, a value subsequently adopted as a parameter in the application of K-Means.

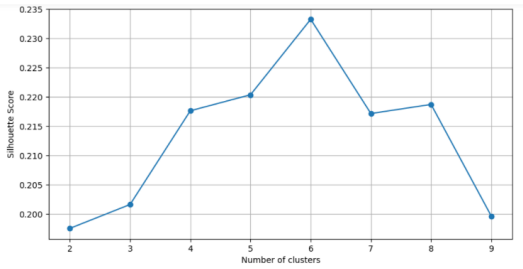


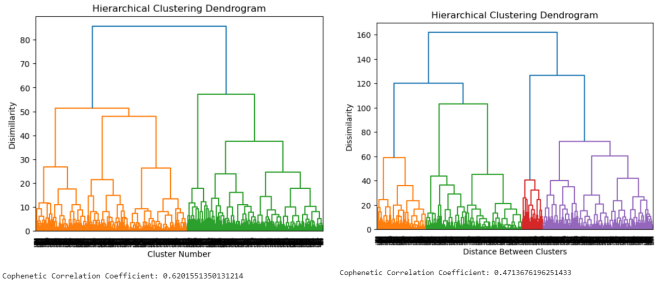
Fig. 4: Silhouette graph showing optimum cluster number

In the training process the primary parameter was the cluster centroid, representing vectors that summarise the average characteristics of the individuals assigned to each cluster. Optimal convergence of centroids is dependent on tuning the cluster number hyperparameter as discussed above.

To ensure reproducibility in the analysis the random seed was hard-coded to a predefined value. As discussed in research by Testa (n.d.), this practice serves to mitigate overfitting by providing consistent results, ensuring each cluster's significance can be interpreted. Nevertheless, the potential for underfitting arises when employing a hard-coded random seed, as the risk of clusters converging to a local minimum is introduced, potentially causing a global optimum to be overlooked.

When employing hierarchical clustering, the primary parameter is the distance metric, which influences the cluster number parameter. Distance metrics measure the dissimilarity between data points, and greatly influence the shape of the produced hierarchies. The most appropriate candidate values included Euclidean, Manhattan, Minkowski and Gower distances among others, with each one measuring distances between data points differently, providing unique advantages and disadvantages.

Cophenetic correlation coefficient was used to measure the efficacy of these candidates, representing how effectively the pairwise distances between the original data points were preserved. Initially, Euclidean distance was employed which yielded identifiable clusters with branches merging on roughly similar levels. Gower distance was then trialled, which, as described by Vagni et al., (2021), is effective with mixed data, such as the sociodemographic data included in the dataset.



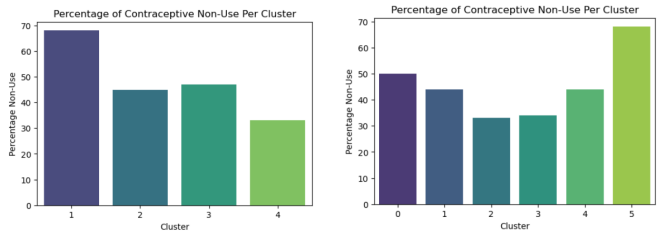
(a) Euclidean-Distance Method (b) Gower-Distance Method

Fig. 5: Dendrograms showing results of linkage matrices for each method

However, the dendrogram for Euclidean distance (fig. 5a) was much more clear-cut and had a higher Cophentic coefficient than for Gower distance (fig. 5b), showing four distinct groups. The dendrogram was consequently clustered into those four groups, allowing for further analysis and comparison with the K-Means clustering method.

3.3 Model Comparison

Comparing the models, it can be observed that the K-Means model (fig. 6b) has means that are more evenly distributed, and with a greater spread, than the Hierarchial model (fig. 6a). This suggests that more distinct communities have been identified, which will allow for easier interpretation of the dataset for policymaking.



(a) Hierarchical Clustering (b) K-Means Clustering

Fig. 6: Bar charts depicting distribution of contraceptive non-use in the target variable amongst the two selected models

Additionally, the Calinski-Harabasz index for hierarchical clustering consistently produced a lower score for a range of seeds, with the K-Means clustering always coming out on top. This suggests that the simpler K-Means result was not only more effective due to better-defined clusters with higher scores, but also more reliable when making policy recommendations based on the discoveries made from clusters, which is perhaps in part due to the interplay between the high quantity of sociodemographic features and the complexity of the correlation between them.

4 Reflection

4.1 Takeaways From Lessons and Coursework

The biggest takeaway learnt in lectures and practised in the coursework was the level of rigour needed in carrying out the methodology behind the machine learning workflow. To perform a successful clustering of a dataset, I learnt from the coursework that many iterations of models are needed, with each iteration revealing a new piece of information about how the hyperparameters should be tuned. I also learnt from the Dimensionality Reduction lecture that when performing PCA, as the number of features increases, the risk of overfitting increases with it. Consequently, if I were to undertake the project again, I would remove the two binary features to help prevent overfitting in the data.

4.2 Module Challenges

One of the most challenging parts of the module for me was employing the mathematical theory behind the machine learning algorithms. There were many libraries available that tended to be 'almost as good' as a custom-made algorithm, oftentimes encouraging me to choose them over designing my own function in the interest of time. This perhaps gave me less control over the training process overall, potentially leading to less useful results. I also found matching a specific problem with a machine learning algorithm difficult, as I discovered that there were always many different approaches to a particular problem that would each produce meaningful results in their own way. Selecting an algorithm required scrutinisation of exactly what it was I wanted to do with the data and needed lots of prior planning before making a start on the programming.

4.3 Next Time

Next time, I would more carefully scrutinise the correlation between variables prior to applying data construction steps. For example, two features in the dataset, the husband's and wife's education, correlate very strongly with the standard of living, potentially allowing these features to be reduced to a single feature. This would allow for more accurate clustering to take place with fewer variables. I also would have taken categorical variables as one-hot encoded instead. With the numerical encoding method used, the potential for feature bias was present in K-Means, with higher-value features for an individual being assigned a greater weight. One-hot encoding would have prevented this issue by assigning each feature a binary value, thereby increasing model performance.

4.4 Project Differentiation

One of the things that differentiates my project is the inclusion of multiple varieties of data representing aspects of a couple's lifestyle like religion and finance. Putting all these pieces together to form an overall picture of an impoverished region sets data scientists up to be able to make predictions with clusters using supervised learning techniques, to target specific communities as opposed to individuals. The project also tackled an imbalanced distribution of the target variable by examining the data's real-world meaning and using both human interpretation and mathematical metrics to put complicated real-world sociodemographic data into context.

References

[1] Brito, C., et al. (2015). Postpartum depression among women with unintended pregnancy. *Revista de Saúde Pública*. doi:10.1590/S0034-8910.2015049005257

[2] Gafar, A., et al. (2020). Determinants of contraceptive use among married women in Indonesia. *F1000Research*, *9*, 280. Retrieved from www.ncbi.nlm.nih.gov/pmc/articles/PMC7137393/

[3] Lim, T. (1997). Contraceptive Method Choice [Data set]. National Family Planning Coordinating Board and the Central Bureau of Statistics. doi:10.24432/C59W2D

[4] Kodinariya, T. M., Makwana, P. (2013). Review on determining the number of clusters in K-Means Clustering. *International Journal of Advance Research in Computer Science and Management Studies, 6*(1), 90-95. Retrieved from https://www.researchgate.net/publication/313554124_Review_on_Determining_of_Cluster_in_K-means_Clustering

[5] Test, L. (n.d.). Assignment 1 - Research and Planning: K-MEANS CLUSTERING. Retrieved from http://18.133.16.176/Files/C++%20CUDA%20MATLAB%20-%20Concurrent%20K-means%20Clustering%20with%20CUDA/Preliminary%20Design.pdf

[6] Ding, C., & He, X. (2004). K-means clustering via principal component analysis. In *Proceedings of the twenty-first international conference on Machine Learning* (p.29). doi:10.1145/1015330.1015408

[7] Patel, V. R., & Mehta, R. G. (2011). Impact of outlier removal and normalization approach in the modified k-means clustering algorithm. *International Journal of Computer Science Issues (IJCSI), 8*(5), 331-336. Retrieved from https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=d316cfbc349d3e7e5b1d6e510b94a3765fe60cc5#page=354

[8] Xu, W., & Xu, Y. (2019). An improved index for clustering validation based on Silhouette index and Calinski-Harabasz index. *IOP Conference Series, 569*(5), 052024. doi:10.1088/1757-899x/569/5/052024