Unsupervised Learning — Final Project

Roni shine Zilberberg , Bar Trabulski

April 2025

Abstract

In this project, we apply several clustering algorithms on the Fetal Health Data. We found the most appropriate clustering algorithm using unsupervised methods. We additionally found the anomalies in the data and removed it. Afterwards, we found the 4 significant features effects on the two dimensions of PCA. Finally, we propose visualization of K-means with PCA algorithm into 2 dimensions with reference to the 4 significant features and visualization of K-means with PCA algorithm into 2 dimensions.The aim of the study was to discover what characterizes sick fetuses and find indicators to identify them. The code is available at https://github.com/Roni_Zilberberg_and_Bar_Trabulski_/Unsupervised-learning.git.

1 Introduction

Unsupervised learning is a field that uses machine learning algorithms to analyze and group unlabeled datasets. These algorithms discover hidden patterns or groupings of data without the need for human intervention [2]. During the project, we used three main methods from this field: clustering, dimensionality reduction, and anomaly detection. Clustering is the task of dividing the population or data points into a number of groups such that data points in the same group have more in common than those in other groups. Simply put, the aim is to segregate groups with similar traits and assign them into clusters [3]. Dimensionality reduction is the process of compressing high-dimensional complex data to a low-dimensional, simpler form preserving its overall structure. High-dimensional data compression is required because high-dimensional data is hard to manage: the data gets sparse because of the curse of dimensionality, and analysis is difficult to process and inefficient. Dimension reduction into 2 or 3 dimensions allows to visualize the structure of the data as well as better connections [4]. Anomaly detection is an important step in the data mining process that aims to identify points of data, events, or observations that significantly deviate from regular patterns in a dataset. In the past decade, machine learning methods have been widely employed to simplify and optimize the efficiency of anomaly detection processes [5]. The aim of the study was to discover what characterizes sick fetuses and find indicators to identify them. In this project, we separate the fetal health labeling feature from the data. Then, we reduce the dimensions by the PCA algorithm because of the high dimensionality of the data. We further apply different clustering algorithms and evaluate their performance by both statistical methods (One-way ANOVA [7] and Kruskal-Wallis [8]) and non-statistical methods (Silhouette, Calinski–Harabasz [14], and inverted Davies–Bouldin [13]). We also relate the external variable to the clusters. Finally, we suggest a visualization of the optimal number of clusters and dimensions combination for the chosen best algorithm.

2 Methods

2.1 Fetal Health Data

The dataset contains 20 numerical attributes, 1 categorical attribute called `histogram tendency` and 1 categorical label attribute called `fetal health` that has three categories: 1 - Normal, 2 - Suspect and 3 - Pathological. The dataset contains 2126 rows in total.[1]

Data preprocessing since the project is about unsupervised learning, we removed the label attribute `fetal health` for the unsupervised learning part and restored it for the **results** section using the pandas library [10].

2.2 PCA (Dimensionality Reduction)

We performed Principal Component Analysis (PCA) on our high-dimensional numerical dataset as an initial step to reduce dimensionality.[6]. This is a linear method that derives new orthogonal components from the features in the dataset, based on the amount of variance they capture, and orders them accordingly. This procedure aimed at enhancing the results of clustering by capturing the critical directions of the data.

2.3 Statistical tests

To test whether there were significant differences between all algorithms performances, we used a *one-way* ANOVA test [7] and Kruskal-Wallis test. [8].
(It is important to note that DBSCAN was excluded from these statistical tests. The algorithm often failed to produce valid cluster assignments, frequently returning NaN scores. Because of that instability, we decided that DBSCAN was not a viable candidate for identifying the optimal clustering model in our analysis).

2.4 Finding the optimal number of clusters

In order to find the most appropriate combination of number of clusters and the dimension of the PCA reduction, we applied a unsupervised grid search running on different number of clusters (between 2 to 10), and on different number of dimensions (also 2 to 10, and without PCA). We used four clustering algorithms: K-means, GMM, Hierarchical, and DBSCAN.

- **K-means Loss** This method is unique for K-means algorithm [11]. For each combination of number of clusters and dimensions, We have calculated the loss and created a heatmap of these values. We found the number of clusters (in 2 dimensions of PCA) from which onwards we got less than 50% of the loss using elbow function of the number of clusters vs the loss value for 2 dimensions.
- **Silhouette score** We used this generic measure for all the clustering algorithms [12]. Although we are aware there is a bias towards low number of clusters, still, we decided choose the best combination according to the highest silhouette score in the heatmap combined with the K-means Loss score for the K-means algorithm that which prevented us from choosing too low a number of clusters despite equal results in the silhouette score.

2.5 Anomaly Detection

In order to detect anomalies in the data, we applied three methods:

- **K-means:** The data was clustered using the K-means algorithm. Then, we defined significant distance from the centroid of a cluster as a higher distance than mean(scores) + $3 \cdot$ std(scores), where std is the standard deviation and score is the distance of a representative form its centroid. point that has a significant distance is suspected to be an anomaly.
- **GMM:** To begin with, the data was clustered according to the GMM algorithm. Each data point was given a log likelihood [15]. The higher the score of the point, the more the point will be considered a non-anomaly, and the lower the score, the more it will be considered an anomaly. We set a cutoff to determine whether a point is an anomaly. The cutoff was set to be 3 std from the mean of the log likelihood. A point that received a score lower than the cutoff is considered an anomaly.
- **One class SVM:** In this method, a hypersphere is constructed so that most of the data points are within it and points outside the margins are considered anomalies [16]. Each point receives a score - the distance of the point from the center of the hypersphere. The higher the score of the point, the more likely it is that the point is an anomaly. We set a cutoff below which only 1 % of the samples are found. Any point whose SVM score is less than the cutoff is considered an anomaly.
- **Isolation Forest:** The data was estimated according to Isolation Forest algorithm, which isolates samples from the data by randomly selecting a feature and a split value between the minimum and maximum of that feature [9]. Anomalies are more likely to be isolated quickly, requiring fewer random splits. For each data point, assigned an anomaly score based on the average path length across all isolation trees. Shorter average paths indicate that the point is most likely an anomaly. We set for the algorithm an expected value of 1% of the data points to identify as anomalies.

2.6 Best Clustering Algorithm

we used Calinski-Harabasz [14] and Davies-Bouldin [13] in order to serve as an additional check on the statistical tests results for finding best clustering algorithm. These methods are internal evaluation scores used to estimate clustering quality without relying on external labels.

3 Results

3.1 Optimal Number of Clusters

First, for each algorithm we chose the best number of clusters. Our data (without the label column) is entirely numeric and because the other algorithms (T-SNE, ICA, ISOMAP) gave similar silhouette scores (~0.42 compared to ~0.41 for PCA) we chose to perform dimensionality reduction with the PCA algorithm (see Methods). We measured the clustering quality using silhouette score for all algorithms by optimizing simultaneously the dimension as well as the number of clusters by unsupervised grid search. In particular, we calculated the loss function values for the K-means algorithm. The best dimension for all the algorithms is 2 with the lowest K-means loss (Fig. 1A) and the highest silhouette score (Fig. 1C,D,E and F). Also, we checked all these methods on the original data without dimension reduction; however, the results were worse. Thus, we decided not to include the results without dimension reduction in the heatmaps.

In order to determine the optimal number of clusters, we chose the highest score of the silhouette values based on the heatmaps, or using the elbow function on the best dimension for the K-means loss as explained in Methods. For K-means the best number of clusters is 7 (combination of less than 50 % loss and highest silhouette—deselected 3 clusters option Fig. 1A,B,C) and for DBSCAN algorithms is 3 (Fig. 1E), compared to hierarchical and GMM with 2 optimal clusters (Fig. 1D,F).

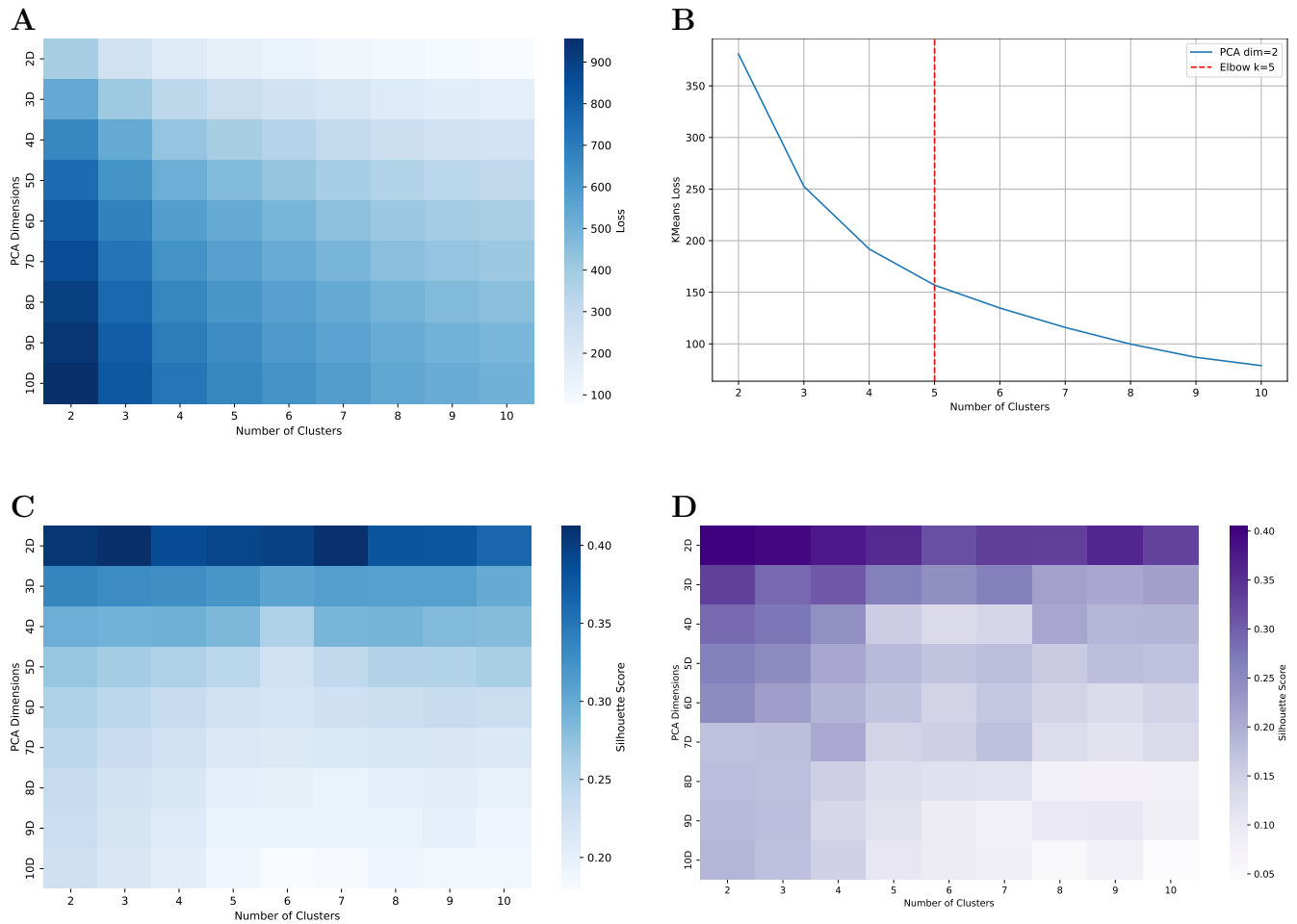


Figure 1: Optimal number of clusters. part 1: K-means loss (A), Elbow graph (B), Silhouette heatmap for K-means (C), Silhouette heatmap for GMM (D).

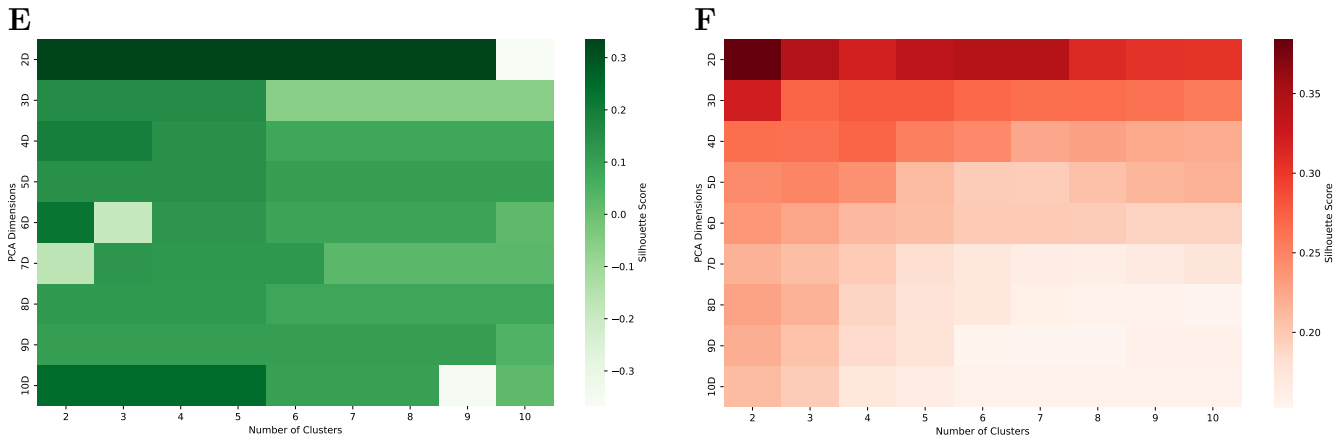


Figure 1: part 2- Silhouette heatmap for DBSCAN (E) and Silhouette heatmap for hierarchical (F)).

3.2 Best Clustering Algorithm

We used the external variable `fetal health` in the one-way ANOVA and Kruskal–Wallis tests with the optimal dimension–cluster combinations for GMM, K-means, and hierarchical clustering. (DBSCAN was excluded due to instability.) According to the silhouette measure, the one-way ANOVA yields a highly significant result $p < 5.58 \times 10^{-128}$. K-means is the best algorithm, with an average silhouette of 0.41, significantly outperforming hierarchical clustering ($p < 2.91 \times 10^{-01}$) and GMM ($p < 5.77 \times 10^{-04}$).

	ANOVA p-value	Kruskal p-value
KMeans	5.58e-128	1.01e-107
GMM	5.77e-04	5.27e-10
Hierarchical	2.91e-01	4.56e-02

Table 1: ANOVA and Kruskal p-values tests were performed for each chosen clustering algorithm separately with reference to the external variable (fetal health).

We then computed a composite score by normalizing and averaging silhouette, Calinski–Harabasz, and inverted Davies–Bouldin indices (higher CH and silhouette are better, lower DB is better), which again crowns K-means as top.

	Silhouette	CH	DB	Composite Score
KMeans	0.41	1665.388	0.857	0.992
GMM	0.41	1706.981	0.985	0.749
Hierarchical	0.38	1568.506	1.027	0.615

Table 2: Silhouette, DB, CH, and Composite scores were performed for each chosen clustering algorithm separately.

As you can see, according to both metrics, we have obtained that the ideal algorithm with the lowest p-value score for both ANOVA and Kruskal–Wallis, as well as the highest composite score of 0.992 is K-means. Therefore, it is the ideal clustering algorithm for our data with two dimensions of PCA and 7 clusters.

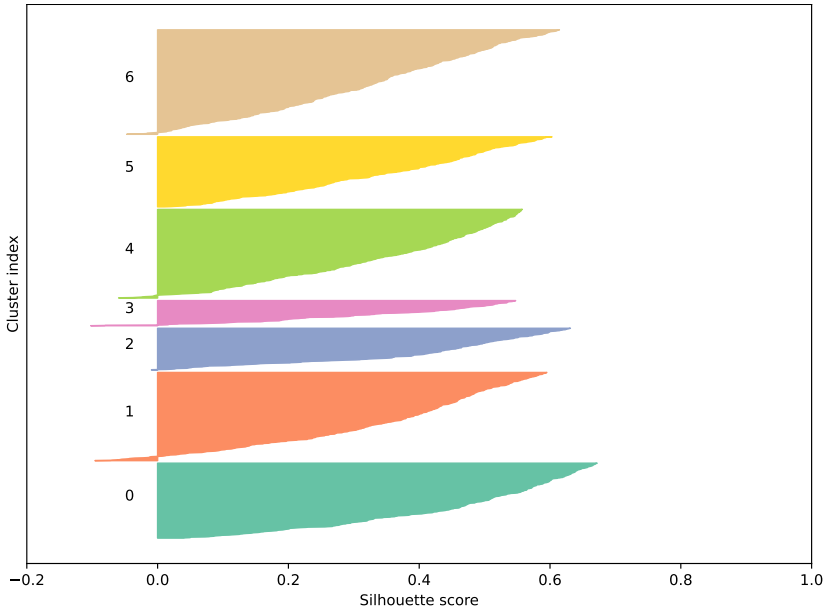


Figure 2: Silhouette graph for K-means algorithm.

3.3 Anomaly Detection

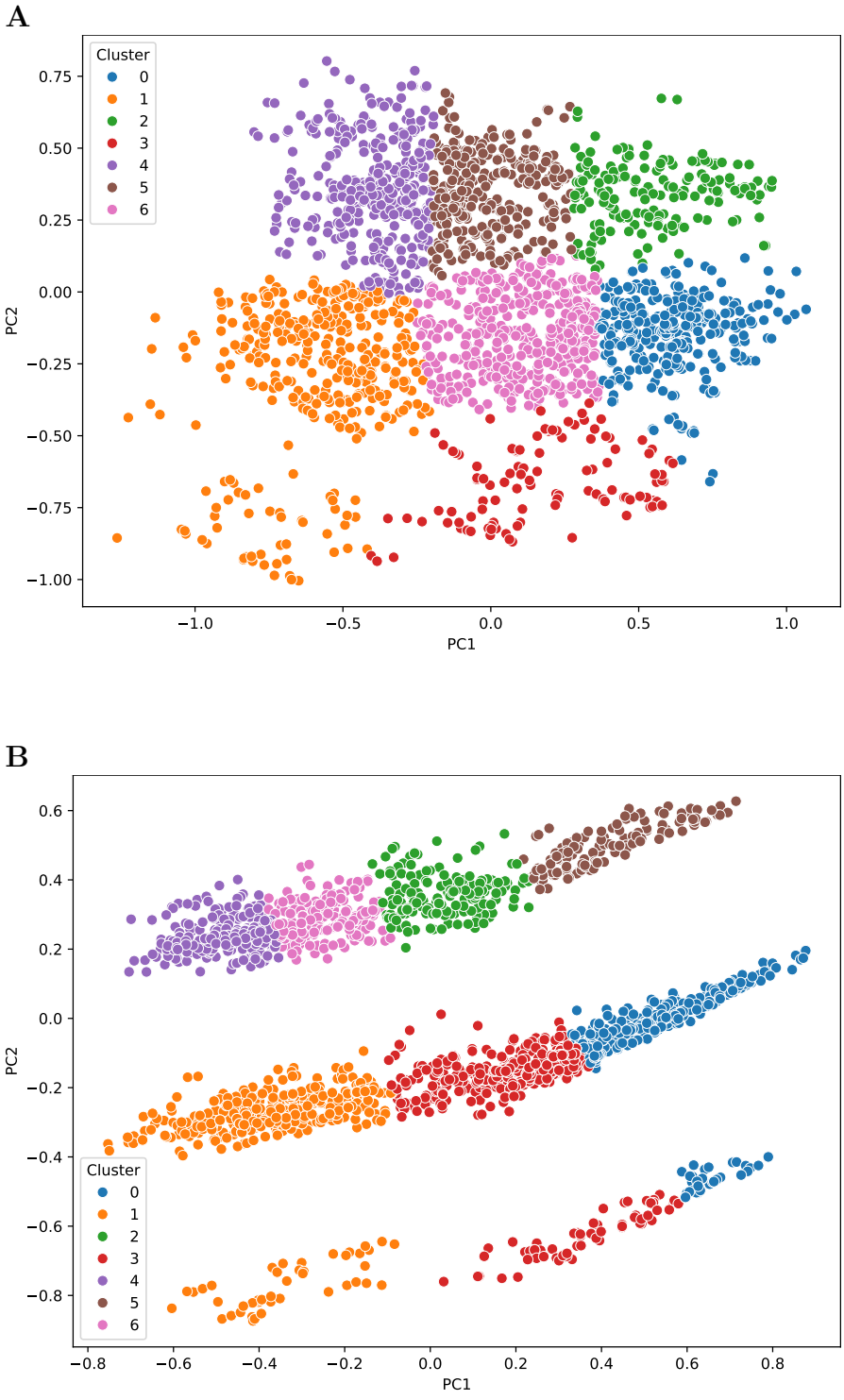
We performed the following four measures to detect anomalies:

- **K-means:** The data was clustered using the K-means algorithm. Then, we defined significant distance from the centroid of a cluster as a higher distance than $\text{mean}(\text{scores}) + 3 \cdot \text{std}(\text{scores})$, where std is the standard deviation and score is the distance of a representative form its centroid.
- **GMM:** The data was clustered according to the GMM algorithm. Each data point was given a log likelihood and we set a cutoff to determine whether a point is an anomaly. The cutoff was set to be 3 std from the mean of the log likelihood.
- **One class SVM:** Each point received a score - the distance of the point from the center of the hypersphere. We set a cutoff below which only 1 % of the samples are found. Any point whose SVM score is less than the cutoff is considered an anomaly.
- **Isolation Forest:** The data was estimated according to Isolation Forest algorithm, each data point, assigned an anomaly score based on the average path length across all isolation trees. We set for the algorithm an expected value of 1% of the data points to identify as anomalies.
- **Anomaly filtering - at least 2 out of 4:** For each of the methods, we created a new feature where a point marked with 1 is an anomaly. Finally, we determined that a point would be considered an anomaly if at least two of the four methods mentioned above determined that it was an anomaly. We obtained that 1.27% of the data was classified as anomalies and decided to removed them because it was a small amount of points (less than 2% of the data).

3.4 Visualization

According to the tests we conducted, we discovered that the `histogram tendency` feature is a strong indicator for identifying sick fetuses.
We found that cluster 3 contains sick fetuses, and after calculating the effects of the variables on the PCA components (Fig. 3C and Fig. 3D), we observed that the higher a cluster appears in the dimensionality reduction graph following K-means clustering (Fig. 3A), the more it is associated with histogram tendency values of 1. Conversely, lower clusters tend to have -1 values, indicating that fetuses in those clusters spend too much time at low heart rates, which may signal distress. Thus, cluster 3 was characterized as pathological.
When focusing on the four features most influential to PCA (severe decelerations, histogram mode, prolonged decelerations, and histogram mean), we observed that cluster 1 contained the majority of sick fetuses (65%) (Fig. 3B).

Moreover, the directions and magnitudes discovered in these four features—the skewness of the heart rate distribution, the width of the histogram, the basal heart rate, and the minimum heart rate value—were useful indicators to distinguish between healthy and sick fetuses across clusters 2, 4, 5, and 6 (Fig. 3A).
We also noticed that the histogram tendency index itself is a dominant factor for determining fetal health: upper clusters in the plot (Fig. 3D) mainly correspond to healthy fetuses, while lower clusters are more associated with sick or suspicious cases.
Finally, surprisingly, we found that the `fetal movement` feature had almost no impact on the PCA components. We would have expected that fetal movement would be a key indicator of fetal well-being, but our results show that heart rate features were much more influential.



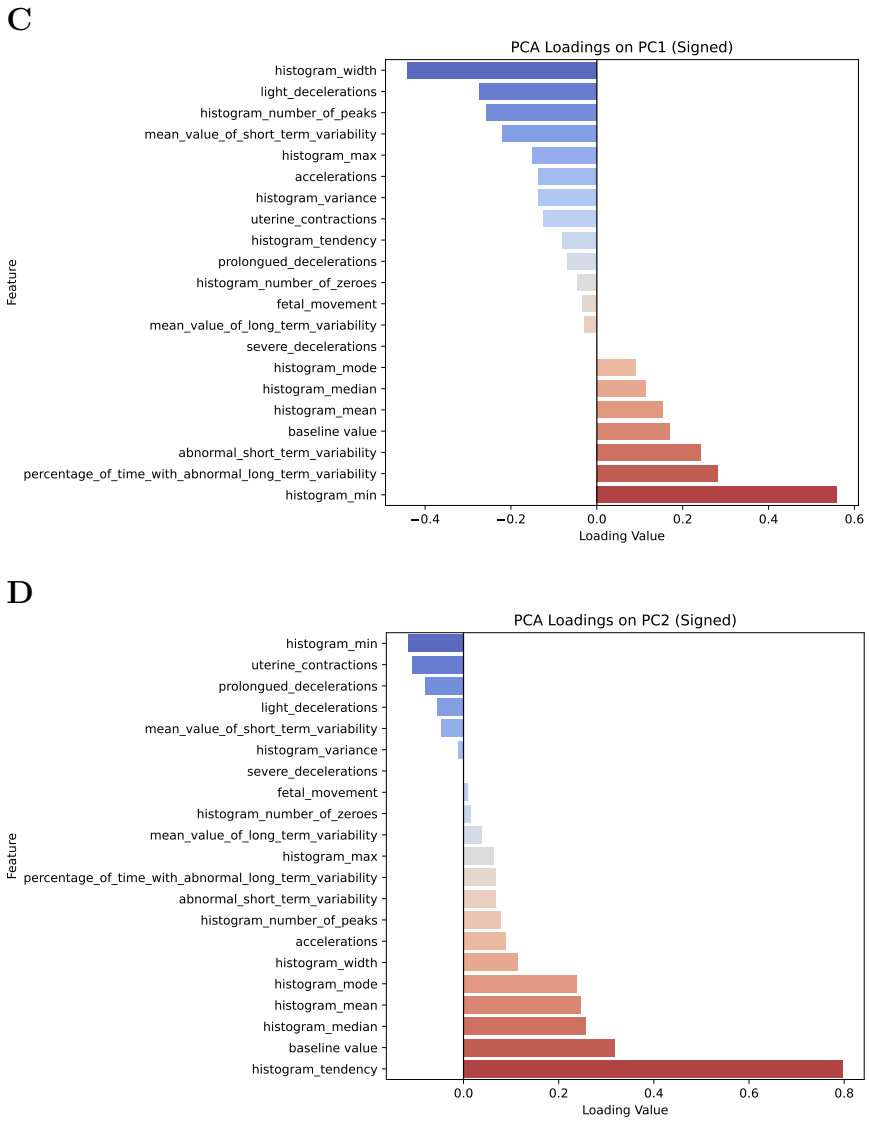


Figure 3: Visualization of K-means with PCA algorithm into 2 dimensions (A), Visualization of K-means with PCA algorithm into 2 dimensions with reference to the 4 significant features of the PCA dimensions (B), directional effects of features on the first dimension of PCA (C), directional effects of features on the second dimension of PCA (D).

4 Discussion

In this project we implemented unsupervised algorithms on the Fetal Health dataset. We found what is the optimal combination of dimensions and number of clusters. All clustering algorithms got 2 dimensions, K-means had 7 clusters (using both on silhouette and K-means loss), DBSCAN had 3 clusters and both hierarchical and GMM had 2 clusters. In addition, we measured the quality of the clusters by performing two approaches, the first, One-way ANOVA and Kruskal-Wallis p-value statistical tests using the external variable fetal health. The second, we computed a composite score by normalizing and averaging silhouette, Calinski–Harabasz, and inverted Davies–Bouldin indices (higher CH and silhouette are better, lower DB is better). Both approaches determined that K-means algorithm is the best. Moreover, we detected the anomalies in the data by using K-means, GMM, one class SVM and Isolation Forest algorithms. Finally, we proposed a nice visualization of 2 dimensions PCA with 7 K-means clusters and also one focusing on the 4 significant features of 2 dimensions PCA components.

In our opinion, fetal movement feature had very little impact, since the dataset did not include duration of each pregnancy. We would like to investigate which features have the most impact on dimensionality reduction given this information.

In addition, we were unable to detect clusters containing a majority of suspect fetuses because their results were not sufficiently unusual. We would like to explore a dataset of these fetuses and examine what makes

them unique to understand why we failed to detect them in our study.
Finally, we believe adding an explainable layer to our composite-score ranking would make the results more practical for clinicians. We suggest training a light model to predict the composite score from the PCA loadings directly and then using SHAP to obtain concise rules (e.g., Severe decelerations > 0.35 $\rightarrow$ pathological). This would enable us to show why cluster 3 is considered to be high-risk, rather than presenting a single silhouette number, and would bridge the gap between unsupervised measures and actual decision support.

References

[1] *Fetal Health Classification*. Kaggle. https://www.kaggle.com/datasets/andrewmvd/fetal-health-classification

[2] *Unsupervised learning*. https://www.ajodo.org/article/S0889-5406(23)00193-2/fulltext

[3] *clustering*. https://ieeexplore.ieee.org/abstract/document/6558109?casa_token=ArqO7D6lPsgAAAAA:1BSAiB2uNvHz7w5h9Z8JJqhtRpVrUmDQBqkyukBAaO6La-aboFqtVrBL4Ybml1B_aDbX0fhu27I

[4] *Dimensionality Reduction*. https://www.scirp.org/journal/paperinformation?paperid=111638

[5] *Anomaly Detection*. https://ieeexplore.ieee.org/abstract/document/4138201casa_token=LGAxetiUIqYAAAAA:5AETl070saOgeA8LLafh5-U0erAJaE6yn-k7wAy-9XdNM8s7DCOBInPYeJltiNix0Q0Pq8ck6Y

[6] *PCA*. https://doi.org/10.1098/rsta.2015.0202

[7] *ANOVA*. https://link.springer.com/chapter/10.1007/978-1-4614-3725-3_8

[8] *Kruskal-Wallis*. https://link.springer.com/chapter/10.1007/978-3-319-30634-6_6

[9] *Isolation Forest*. https://ieeexplore.ieee.org/abstract/document/10108034

[10] *pandas*. https://d1wqtxts1xzle7.cloudfront.net/117768488/pyhpc2011_submission_9-libre.pdf?1724818056=&response-content-disposition=inline%3B+filename%3DAnaplastic_Lymphoma_Kinase_ALK_positive.pdf&Expires=1745716844&Signature=UiTtpcyDLhEW-620GsGQ9qp1IHmOrK-Xx7fhW8-8-xm1zARmoGEr4jtFf-5SIP0q1y7I7N~bG4OzZ89Vw5aeGItiUWsTl_&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA

[11] *Loss function* https://link.springer.com/article/10.1007/s40745-020-00253-5

[12] *Silhouette score* https://ieeexplore.ieee.org/abstract/document/9260048?casa_token=QMj2RGy52S0AAAAA:XpjuJeIbIaDgSMBdJ0_qu6AeDUO0mYod-mqQA1aqoXGGPx4mPIkDhtxmOS7k3ATUPzDj0Sd-zY0

[13] *Davies-Bouldin* https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=b2db00f73fc6b97ebe12e97cfdaefbb2fefc253b

[14] *Calinski–Harabasz* https://jurnal.polibatam.ac.id/index.php/JAIC/article/view/4947

[15] *Likelihood* https://link.springer.com/chapter/10.1007/978-1-349-20865-4_16

[16] *One-class SVM* https://www.sciencedirect.com/science/article/pii/S0360835205000100?casa_token=MJcFcKI5UK0AAAAA:EzlP5x3AjxhENZzEW5R1zzfD2Mx4kaIAsnsYwZi1vfGbR698545zItF7_shmK1jH1etS32CPlOE