Multidimensional Visualization of Maternal Health Data

Indrė Blagnytė

Vilnius University, Faculty of Mathematics and Informatics,
Naugarduko g. 24, Vilnius, Lithuania
indre.blagnyte@mif.stud.vu.lt

Abstract. Visualizing multidimensional health data poses challenges in selecting methods that effectively reveal patterns and separations. This study evaluates five visualization techniques for maternal health risk data: scatter plot matrix, parallel coordinates, RadViz, principal component analysis (PCA), and multidimensional scaling (MDS). Both standardized and normalized data are used to assess group separation effectiveness. Direct visualization methods and PCA show limited separation, especially for medium-risk. MDS with Manhattan distance and standardized data provides the best separation. Results show that the visualization method determines the ideal scaling approach, with no single technique universally optimal for multivariate health data.

Keywords: Multidimensional Visualization, scaling, direct visualisation, PCA, MDS.

1. Introduction

Data visualisation can be a useful tool for public health specialists and researchers to support decision making; however, choosing the best visualization technique can be tricky [1][2]. Health data tends to be complex, and visualising it in understandable ways has applications not only for diagnostic purposes but also for public communication of health information [3]. Health data tends to be multivariate, however, most common visualisation techniques are only meant for two or three-dimensional data [3][4]. For this specific multidimensional data, visualization techniques are necessary [4][5].

A different obstacle is choosing the best scaling method for health data [6]. A 2021 study by M. Ahsan explored the effect of six different scaling methods on Machine learning models using health data. The scaling methods affected the model accuracy, however, no one method was universally best.

In this study, the data chosen was maternal health risk data that has been collected from hospitals and clinics in rural areas of Bangladesh [7].

DOI: https://doi.org/10.15388/LMITT.2025.2

It was first gathered for the purpose of diabetes research, but later applied for evaluating pregnancy risk. The goal of this analysis is to find the best type of visualization method for showing group separation of this kind of health data. As well as exploring the differences between standardization and normalization for this visualization task, and finding the preferred one.

Five visualization methods were chosen for the visualization task. Three direct visualization methods: scatter plot matrix, parallel coordinates, and RadViz, and two dimensionality reduction techniques: principal component analysis (PCA) and multidimensional scaling (MDS). The scatter plot matrix visualizes all possible pairwise combinations of features as scatter plots [5]. For the parallel coordinates plot, each feature is represented by a parallel axis, and each n-dimensional point is represented by a polyline crossing each of the n axes at the appropriate feature value [5][8]. RadViz generates nonlinear mappings of high-dimensional data onto a plane by modelling a physical spring system where the variables create anchor points [9]. PCA linearly transforms high-dimensional data so that most of the variance is conserved in the first few components, allowing for dimensionality reduction by elimination of the last few components [5][10]. MDS, when used for dimensionality reduction, uses a pairwise distance matrix and tries to find low-dimensional points so that distances between the points in the low-dimensional space are as close to the original proximities in the matrix [5][11]. This uses nonlinear transformations, and the type of distances chosen affects the results of the analysis [11][12].

2. Dataset

The chosen dataset was Maternal Health Risk data from the UCI Machine Learning Repository. The data set has 1013 instances, 7 total features, 6 of which are numerical and 1 categorical. The categorical variable represents predicted risk level during pregnancy, the possible values being: low-risk, mid-risk, and high-risk. The numerical variables are as follows: age, systolic or upper blood pressure, diastolic or lower blood pressure, blood glucose levels, body temperature, and resting heart rate. All analyses were done with both normalized (to range [0,1]) and standardized (mean 0, standard deviation 1) data, and results compared. The visualisation was judged on group separation by the categorical variable, where the best group separation was determined visually based on two criteria: the visible distinctness between clusters representing different risk groups, and the reduced overlap area among groups compared to other methods.

3. Direct visualization and comparison to Random Forest.

Three direct visualization methods were attempted: scatter plot matrix, parallel coordinates plot, and RadViz plot. The standardized and normalized data plots were almost identical; only the standardized data plots were included. Random Forest analysis was also performed to compare the variable importance with the significance of variables in the direct visualization methods. Random Forest showed that blood sugar had the highest importance for predicting risk level, with all others lagging behind, and body temperature showing the lowest importance (Fig. 1).

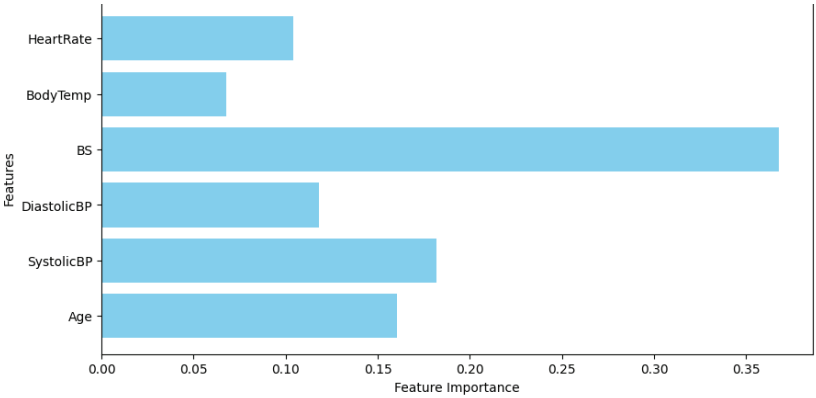

Figure 1. Feature importance using the Random Forest algorithm. The highest importance is shown by blood sugar. Generated using Python 3.11.9.

3.1 Scatter plot matrix

From the individual feature graphs, we can see low-risk and medium-risk groups following a similar distribution and high-risk groups differing from these (Fig. 2). The exceptions to this are blood sugar, where all the distributions differ, and body temperature, where high and medium-risk are more similar. For the pairwise plots, there can be seen some separation between high-risk and low-risk in all pairs with blood sugar, especially in the blood sugar and age pairwise plot. This is consistent with blood sugar showing the highest importance for Random Forest. However, there is still a lot of separation, and the medium-risk group is not separated from the other groups.

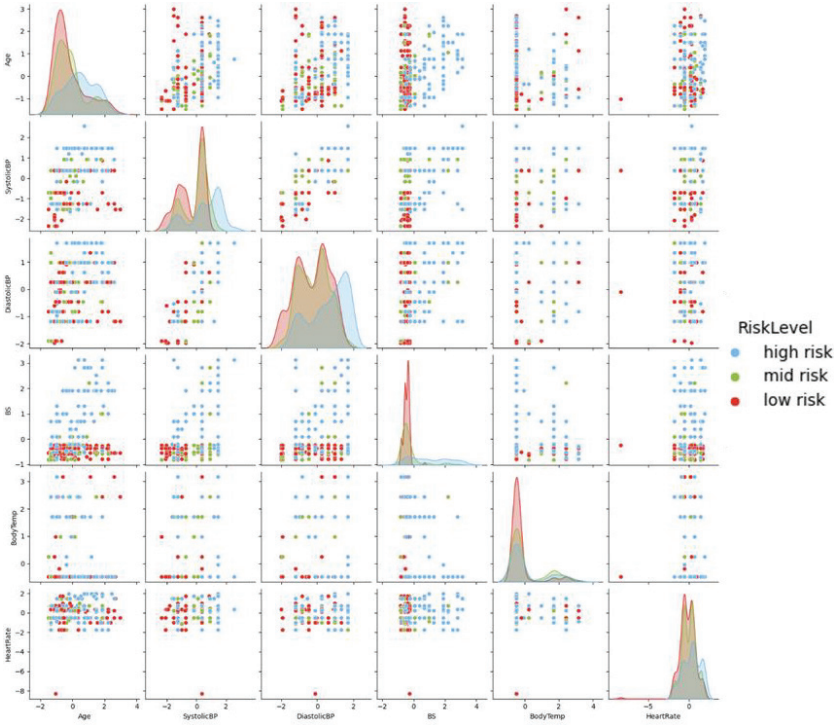

Figure 2. Scatter plot matrix of the standardized Maternal Risk data. The clearest separation between groups can be seen in the blood sugar versus age plot, where high-risk cases tend to form a distinct cluster away from low and medium-risk groups. However, medium and low-risk groups still show significant overlap across most feature combinations. Generated using Python 3.11.9.

3.2 Parallel coordinates plot

The ranges of the groups in the parallel coordinates plot (Fig. 3) overlap significantly, so the groups do not have separation. The mean lines, however, show some separation for all features except body temperature, where the high-risk and mid-risk lines overlap, which might explain the low importance of this variable in Random Forest. Overall, high-risk shows the best separation from the other groups.

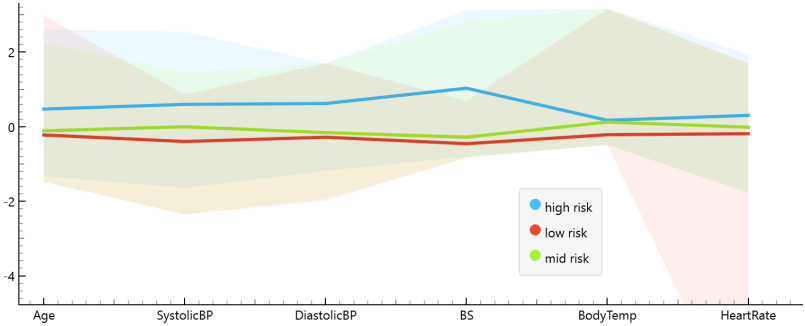

Figure 3. Parallel coordinates plot of the standardized Maternal Health Risk data, with the individual lines hidden and only range and mean shown. The high-risk group's mean line consistently deviates from the low and medium-risk groups, especially for blood sugar and systolic blood pressure. Body temperature, however, shows substantial overlap between all groups. Generated using Orange 3.38.1.

3.3 RadViz plot

RadViz plot (Fig. 4) shows some separation between high-risk and low-risk. However, there is still a good amount of overlap. The medium-risk group overlaps heavily with both high-risk and low-risk. Interestingly, the groups seem to separate out mostly by heart rate, which showed quite small importance in Random Forest, and not blood sugar, which showed the largest importance.

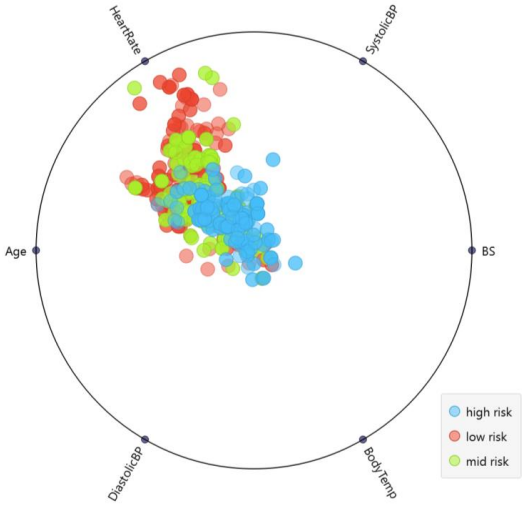

Figure 4. RadViz plot of the standardized Maternal Health Risk data. Some separation can be seen between high-risk and low-risk groups, mainly along the heart rate anchor. Generated using Orange 3.38.1.

4. Principal Component Analysis (PCA)

For the standardized data, the variance explained by the first two principal components was 62.53%, and for normalized data, 70.24%. Both PCA plots showed a lot of overlap (Fig. 5A, B). The most separated group was the high-risk group, with medium and low-risk groups being deeply overlapped. Of the two versions, the normalized data was slightly preferable because of the similar amount of separation and higher explained variance. RadViz plot arguably showed better overall group separation than PCA visualization, though not significantly, and therefore, PCA was not necessarily optimal for visualizing this dataset.

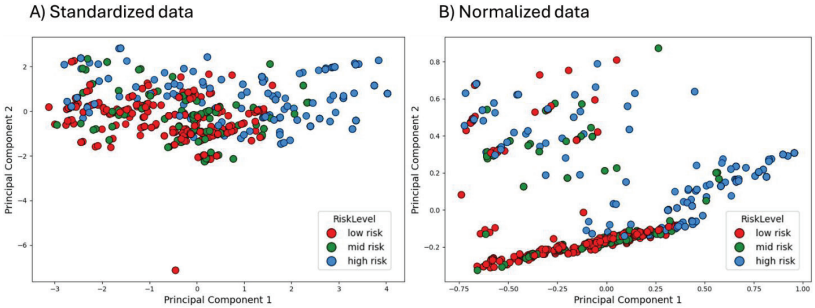

Figure 5. Scatter plot for the first two principal components: A) using standardized data, B) using normalized data. Although the high-risk group forms a somewhat distinct cluster, the medium and low-risk groups overlap heavily along the first two principal components. Generated using Python 3.11.9.

5. Multidimensional Scaling (MDS)

Multidimensional scaling was attempted using standardized and normalized data, using both Euclidean and Manhattan distances. Visualizations using Euclidean distances did not show any better separation than PCA. Using Manhattan distances with the normalized data showed significantly better separation of the high-risk group from others when compared to both PCA and direct visualization methods (Fig. 6A). However, the plots showed poor separation of the low and medium-risk groups. For the standardized data, the separation for the high-risk was not as good, however, the overall separation between all three groups was better (Fig. 6 B). Of all visualization methods, MDS using Manhattan distances and standardized data showed

the best group separation between all groups. Better than PCA and better than all direct visualization methods.

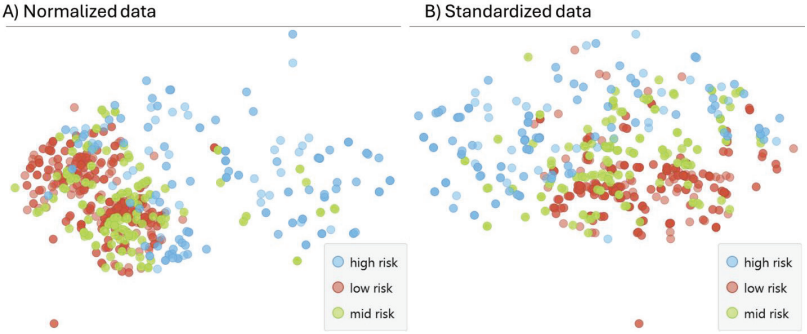

Figure 6. MDS plot using Manhattan distances and PCA initial solution: A) done with normalized data, B) done with standardized data. High-risk instances form a distinct cluster, while medium and low-risk groups overlap. Standardised data show improved low and medium-risk group separation. Generated using Orange 3.38.1.

6. Conclusions

This study explored visualization methods for maternal health risk data and the impact of standardization versus normalization. To objectively assess the quality of group separation, visual inspection was combined with consideration of group compactness and distinctness from other groups. All direct visualisation methods showed limited separation, with medium-risk being most overlapped. Choosing between normalization and standardization did not impact this separation. PCA was also not optimal for showing group separation, with normalization being preferred due to higher explained variance of the first two components. MDS using Manhattan distance and standardized data showed the clearest differentiation between low, medium, and high-risk groups, with less visual overlap than PCA, RadViz, or direct visualization techniques. Thus, based on these qualitative criteria and consistent visual patterns across repetitions, MDS was judged the most effective method. In terms of Normalization vs standardization, neither was preferred for all methods, and therefore, while standardization was optimal for the best method, the conclusion was that the method of visualization dictates the scaling method just as much as the nature of the data itself.

References

[1] Park, S., Bekemeier, B., Flaxman, A., & Schultz, M. (2021). Impact of data visualization on decision-making and its implications for public health practice: a systematic literature review. *Informatics for Health and Social Care*, *47*(2), 175–193.

[2] Austin, R. R., Mathiason, M. A., & Monsen, K. A. (2022). Using data visualization to detect patterns in whole-person health data. *Research in Nursing & Health*, 45, 466–476.

[3] O'connor, S., Waite, M., Duce, D., O'Donnell, A., & Ronquillo, C. (2020). Data visualization in health care: The Florence effect. *Journal of Advanced Nursing*.

[4] Brunsdon, C., Fotheringham, A. S., & Charlton, M. E. (1998). An investigation of methods for visualising highly multivariate datasets. *Case Studies of Visualization in the Social Sciences*, *43*.

[5] Dzemyda, G., Kurasova, O., Žilinskas, J. (2012) Multidimensional Data Visualization: Methods and Applications. Springer Optimization and Its Applications. Springer New York.

[6] Ahsan, M. M., Mahmud, M. A. P., Saha, P. K., Gupta, K. D., & Siddique, Z. (2021). Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance. *Technologies*, *9*(3), 52.

[7] Ahmed, M., Kashem, M.A., Rahman, M., & Khatun, S. (2020). Review and Analysis of Risk Factor of Maternal Health in Remote Area Using the Internet of Things (IoT). Lecture Notes in Electrical Engineering.

[8] Moustafa, R.E. (2011), Parallel coordinate and parallel coordinate density plots. WIREs Comp Stat, 3: 134-148.

[9] Rubio-Sánchez, M., Raya, L., Diaz, F., & Sanchez, A. (2015). A comparative study between radviz and star coordinates. *IEEE transactions on visualization and computer graphics*, *22*(1), 619-628.

[10] Ringnér, M. (2008). What is principal component analysis?. *Nature biotechnology*, *26*(3), 303-304.

[11] Buja, A., Swayne, D. F., Littman, M. L., Dean, N., Hofmann, H., & Chen, L. (2008). Data Visualization With Multidimensional Scaling. *Journal of Computational and Graphical Statistics*, 17(2), 444–472.

[12] Muller, S. T. (2022). *Distance, similarity, and multidimensional scaling*