Data Description and Exploratory Analysis

The dataset used in this analysis is from the UC Irvine Machine Learning Repository and is named Breast Cancer Wisconsin (Diagnostic). It was provided on October 31, 1995, and includes 569 observations with 10 numerical features: radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. Additionally, there is one target variable, Diagnosis, which is categorical with two classes: M (malignant) and B (benign). This dataset provides real-valued measurements that help in studying and predicting breast cancer diagnoses based on various diagnostic features, [1].

The dataset has unbalanced classes, with class M (malignant) being one-third more common than class B (benign). The pair plot of features in the Breast Cancer dataset (Figure 1) shows that the variables radius, perimeter, and area are highly positively correlated. Additionally, concavity, compactness, and concave_points also have strong relationships. The density plots on the diagonal reveal a clear separation between the two classes for several features, especially radius, area, and concavity. Furthermore, the box plots (Figure 2) provide an even clearer view of this separation, highlighting the differences between malignant and benign cases based on these diagnostic measurements.

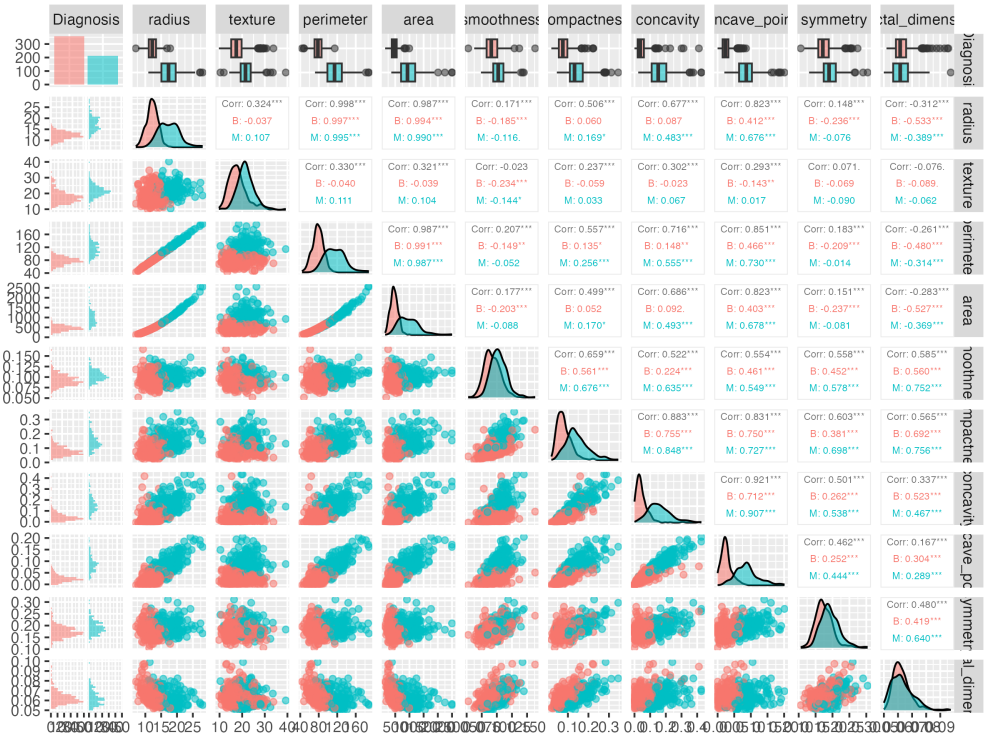

Figure 1: Pair Plot of Features in Breast Cancer Dataset

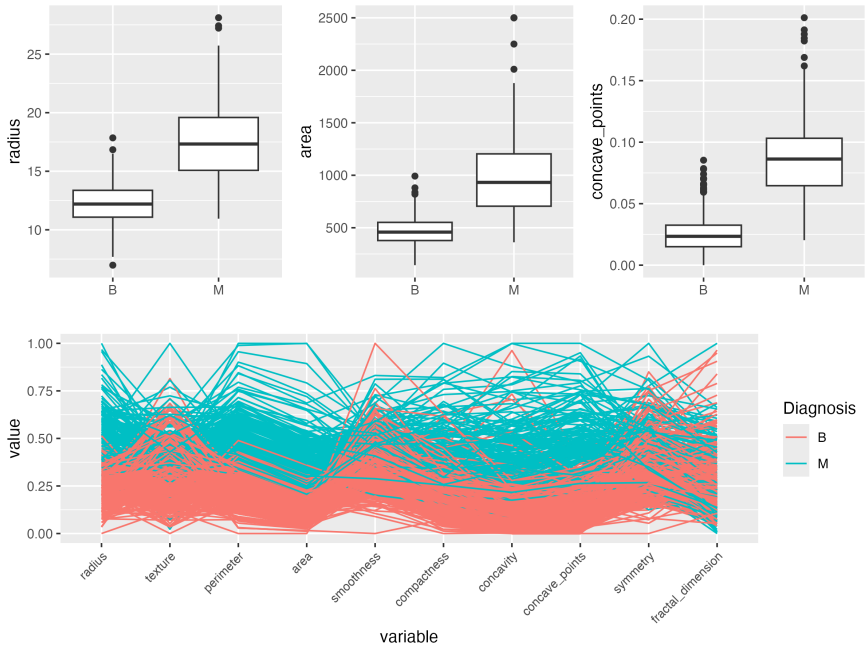

Figure 2: Box Plots and Parallel Coordinates Plot

Problem Statement

Breast cancer diagnosis heavily depends on recognizing patterns in tumor characteristics to differentiate between malignant and benign cases. However, in many real-world scenarios, the available samples often lack labels, and obtaining labeled data can be difficult and impractical. Clustering techniques provide an unsupervised approach to discovering hidden patterns in the data, grouping similar cases without requiring prior class labels.

This analysis applies Mixture Model-Based Clustering to the Breast Cancer Wisconsin dataset, comparing Gaussian Parsimonious Clustering and the Mixture of Factor Analyzers. The goal is to assess which method provides better cluster separation and accuracy in identifying breast cancer cases.

Method

(i) Gaussian Parsimonious Clustering Model (GPCM)
GPCM is used to group similar data points by assuming each group follows a Gaussian distribution. It allows each cluster to have its own shape, size, and orientation by adjusting the spread and direction of the distribution. The model represents each group's spread using the covariance matrix Σ_g, which is decomposed as: $\Sigma_g = \lambda_g D_g A_g D_g'$. The model estimates these group properties using the Expectation Maximization (EM) algorithm, which iteratively refines the group assignments until the best fit is found. Finally, using the Bayesian Information Criterion (BIC) to find the best clustering model by comparing different versions of GPCM.

(ii) Mixture of Factor Analyzers Model (MFA)
The MFA model is a clustering method designed for high-dimensional data. It assumes that each observation is affected by a smaller set of hidden factors. The model represents each data point as: $X_i = \mu_g + \Lambda_g U_{ig} + \epsilon_{ig}$, which Λ_g is a factor loading matrix, U_{ig} represents latent factors. The model

assigns probabilities to each cluster using the density function:

$$f(x|\vartheta) = \sum_{g=1}^{G} \pi_g \phi(x|\mu_g, \Lambda_g \Lambda_g' + \Psi_g)$$

Then, the Alternating Expectation-Conditional Maximization (AECM) algorithm is used to estimate parameters.

To begin with, I applied an unsupervised method, setting the cluster range from 1 to 5. The model determined that the optimal number of clusters was 5, but the Adjusted Rand Index (ARI) was low, at 31.13%. Since the dataset actually contains only two classes, the unsupervised method did not produce a meaningful result. As a result, I decided to switch to a semi-supervised method, where I fixed the number of clusters to 2.

I created three different models for each method, resulting in a total of six models and to compare them with the original class labels. The first model used k-means initialization, the second model used 10 random starts, and the third model used 20 random starts.

Results and Discussion

Overall, the models did not perform very well as expected. The MFA-2 model achieved the lowest misclassification rate (8.96%) and the highest ARI (67.04%), but the GPCM-2 and GPCM-3 models showed slightly better BIC scores. These results indicate that a better BIC score does not always correspond to a better model, and increasing the number of random starts does not necessarily improve performance. In this study, the MFA model performed marginally better than the GPCM models, though it required more computation time. The models' performance may be affected by outliers in many of the attributes, as most variables contain a small number of outliers. Additionally, the relatively small size of the dataset may have contributed to the limited performance. Since some variables are highly correlated, reducing dimensionality might help reduce noise and improve separation between clusters.

Model	Structure	BIC	MCR	ARI	Parameters
GPCM-1	VVV	-3336.50	9.84%	64.30%	K-mean
GPCM-2	VVV	-3318.49	10.90%	61.00%	10 random start
GPCM-3	VVV	-3318.49	10.90%	61.00%	20 random start
MFA-1	UUU	-3347.21	10.90%	61.02%	K-mean, G=2, q=5
MFA-2	UCU	-3890.82	8.96%	67.04%	10 random start, G=2, q=5
MFA-3	UUU	-3746.36	12.12%	57.19%	20 random start, G=2, q=3

Table 1: Comparison of model structures and performance metrics

Conclusion

In summary, both the GPCM and MFA methods provided moderate performance on this dataset, with MFA-2 achieving a lower misclassification rate but requiring more computation time. These result also shows that the influence of dataset size and outliers on clustering performance. Future work could involve addressing outliers, using dimensionality reduction, and testing different evaluation methods or data subsets for further improvements.