Unsupervised Learning for Tumor Segmentation in the Breast Cancer Wisconsin Dataset

Unsupervised Machine Learning Course Final Project

Sin Yee Wu (Mandy)

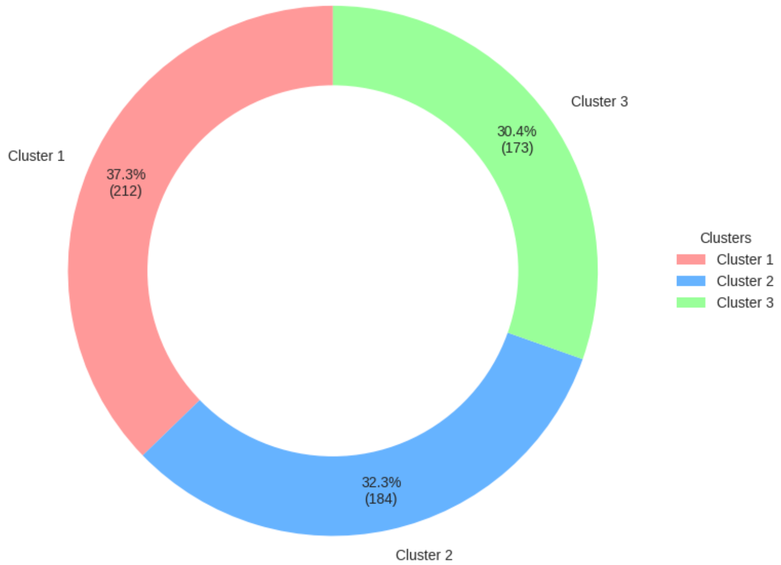
2024-10-21

Table of Contents

1. EXECUTIVE SUMMARY

In this study, I employed unsupervised learning techniques to analyze the Breast Cancer Wisconsin (Diagnostic) Dataset. My goal was to identify distinct groups of breast tumors based on their cellular characteristics, potentially uncovering new insights into tumor classification and improving diagnostic procedures. I used K-Means clustering, Hierarchical Clustering, and DBSCAN, complemented by dimensionality reduction techniques.

The analysis revealed a complex structure within the data, suggesting a spectrum of tumour characteristics rather than clearly delineated categories. The donut chart visualizes the distribution of samples across the three identified clusters:



1. Cluster 1 (light red)
   - 212 samples
   - 37.3% of the total
   - Occupies the largest segment of the donut
2. Cluster 2 (light blue)
   - 184 samples
   - 32.3% of the total
   - Second largest segment of the donut
3. Cluster 3 (light green)
   - 173 samples
   - 30.4% of the total
   - Smallest segment of the donut, but still close in size to the other two

This insight could lead to more nuanced diagnostic approaches and personalized treatment strategies in breast cancer care. The silhouette score of 0.298 (±0.030) indicates a reasonable clustering quality, suggesting that while the clusters are distinguishable, there is some overlap between them.

2. INTRODUCTION

2.1. Main Objective of the Analysis

My main objective in this analysis is to apply clustering techniques to the Breast Cancer Wisconsin (Diagnostic) Dataset. I aim to identify distinct groups of tumors based on their cellular characteristics, potentially uncovering new insights into tumor classification. This unsupervised learning approach could lead to improved diagnostic procedures and more personalized treatment strategies in breast cancer care. By focusing on clustering, I hope to reveal patterns that might not be apparent in the traditional binary (benign/malignant) classification, offering a more nuanced understanding of tumor variability.

Potential benefits to healthcare providers, patients, and medical research:

- Improved tumor classification without relying on predefined categories
- Potential discovery of previously unrecognized tumor subgroups
- Enhanced understanding of the relationships between cellular features and tumor behavior
- Support for more targeted treatment approaches based on tumor characteristics
- Contribution to the broader field of precision medicine in oncology

2.2. Background on Breast Cancer Diagnosis

Breast cancer diagnosis typically involves imaging techniques and tissue analysis, often using Fine Needle Aspiration (FNA). While traditional approaches categorize tumors as benign or malignant, our unsupervised learning techniques aim to reveal more nuanced groupings that may not align perfectly with this binary classification. By applying clustering algorithms to the cellular features extracted from FNA samples, we aim to complement current diagnostic methods by:

1. Identifying potential subtypes within benign and malignant categories
2. Highlighting cases that may require additional scrutiny due to atypical feature combinations
3. Providing a data-driven approach to tumor classification that can adapt to new information

This unsupervised approach can work alongside traditional diagnostic methods, offering additional insights and potentially improving the accuracy and personalization of breast cancer diagnosis and treatment planning.

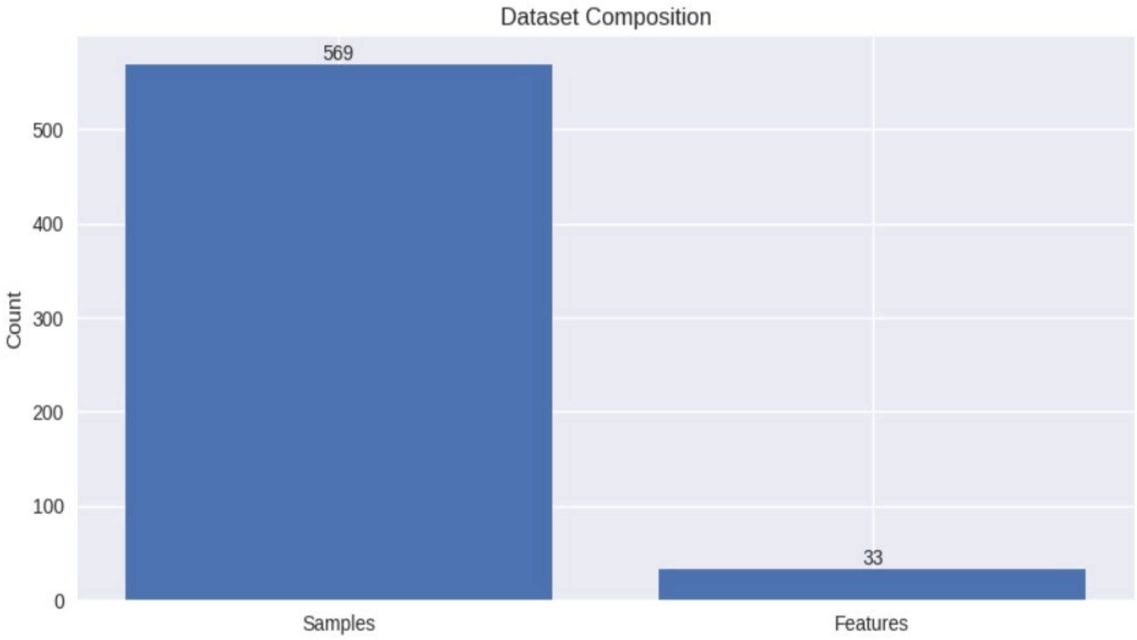
3. DATASET DESCRIPTION

3.1. Overview of the Breast Cancer Wisconsin (Diagnostic) Dataset

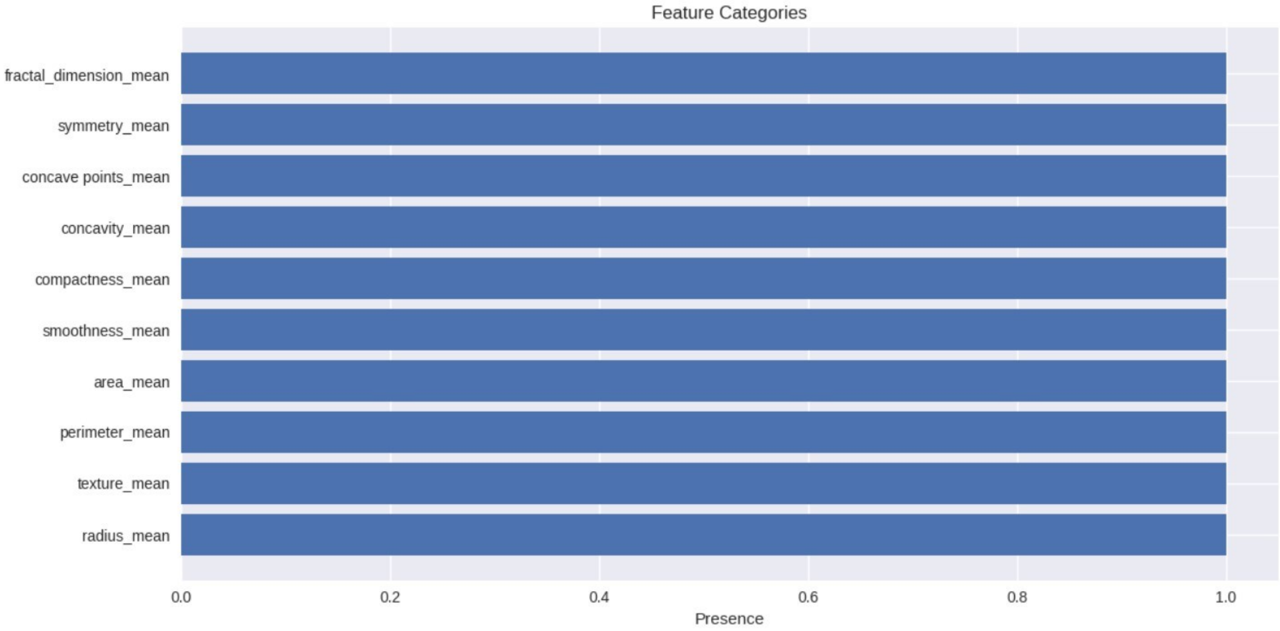
Source: Kaggle Breast Cancer Wisconsin (Diagnostic) Dataset

The Breast Cancer Wisconsin (Diagnostic) Dataset is commonly used for machine learning tasks related to breast cancer diagnosis. The dataset contains features computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. Each row represents a sample, with various measurements of cell nuclei characteristics.

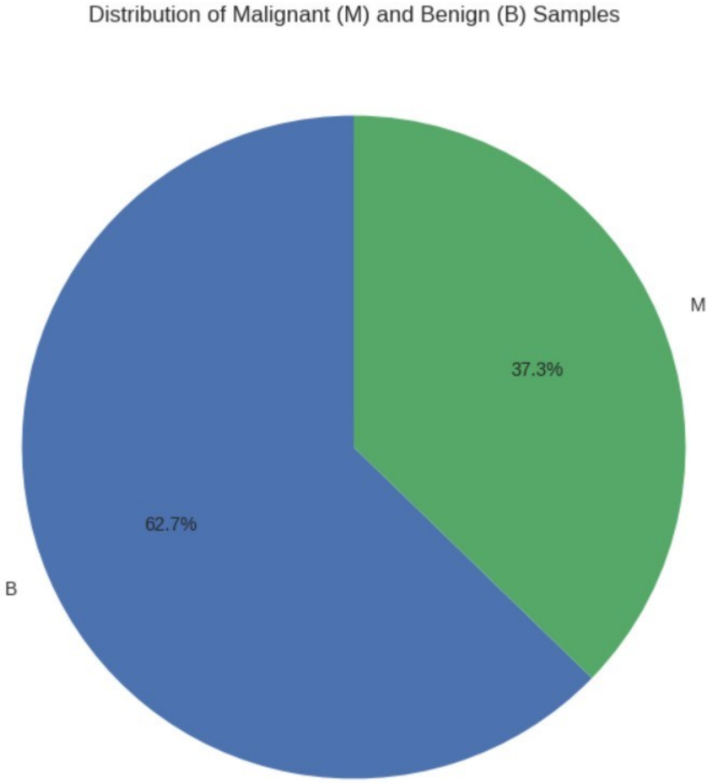
The dataset has 569 samples (rows). There are 32 columns, including an ID column, a diagnosis column (M for malignant, B for benign), and 30 feature columns. The features include measurements like radius, texture, perimeter, area, smoothness, compactness, concavity, concavity, concave points, symmetry, and fractal dimension. For each feature, there are three values: mean, standard error (se), and "worst" or largest (worst).



Bar Chart of Samples and Features shows the dataset contains 569 samples and 30 features, giving a quick overview of the dataset's size and complexity.



Horizontal Bar Chart of Mean Features: displays the average values for each feature across all samples, helping to identify which cellular characteristics tend to have higher or lower measurements overall.



Pie Chart of Diagnosis Distribution: illustrates the proportion of benign and malignant tumours in the dataset, providing context for the balance between the two main categories.

3.2. Relevance to the Analysis

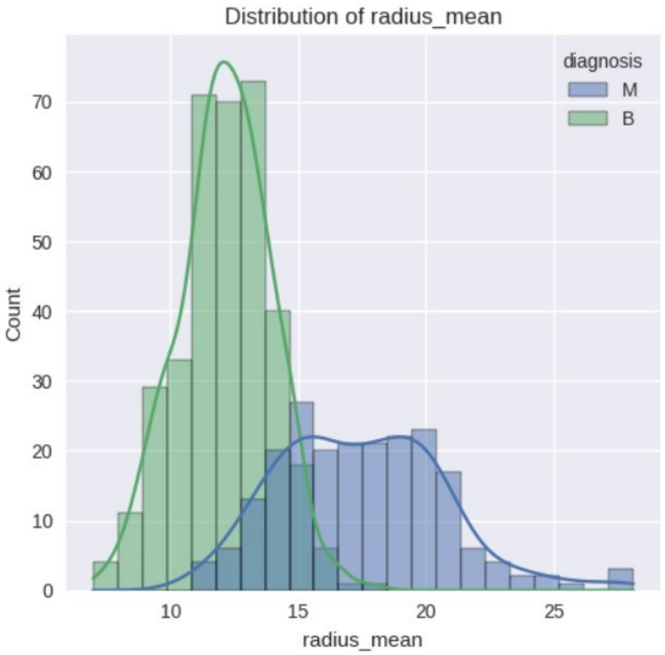
This dataset is well-suited for the unsupervised learning project on tumor segmentation because:

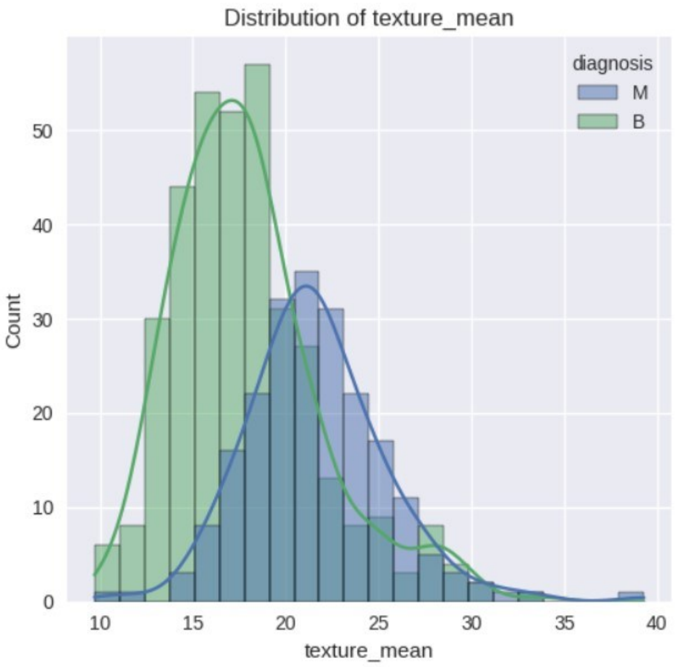
- It contains real-world medical data relevant to breast cancer diagnosis.
- The features provide detailed information about cell nuclei characteristics, which can be used to identify patterns or groups.
- The presence of both benign and malignant samples allows for potential validation of clustering results.
- The number of features and samples is sufficient for meaningful unsupervised learning analysis.

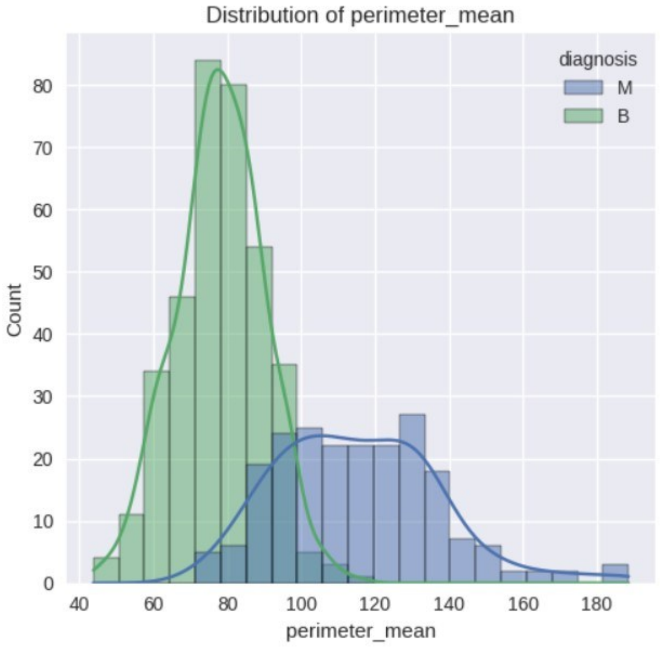
3.3. Feature Significance in Breast Cancer Diagnosis

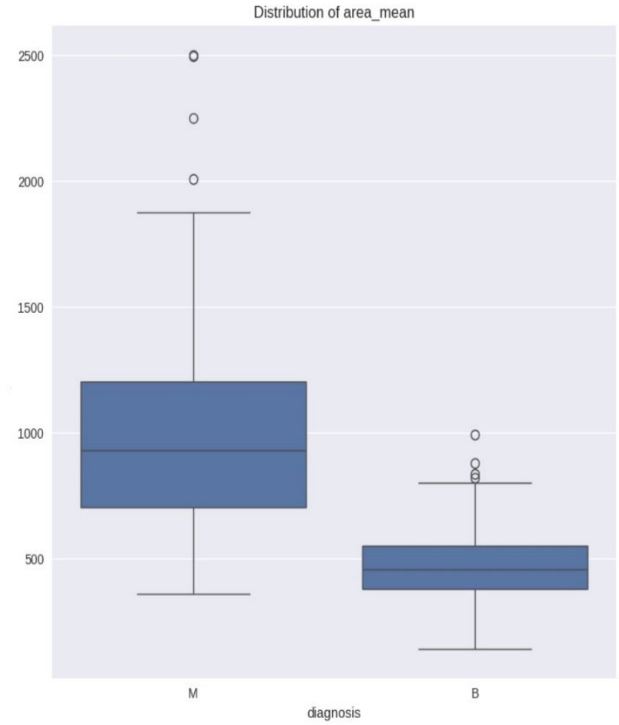
The features in this dataset represent various aspects of cell nucleus morphology. Key features include radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. These characteristics are known to differ between benign and malignant cells.

The distribution of some key features is visualized below:









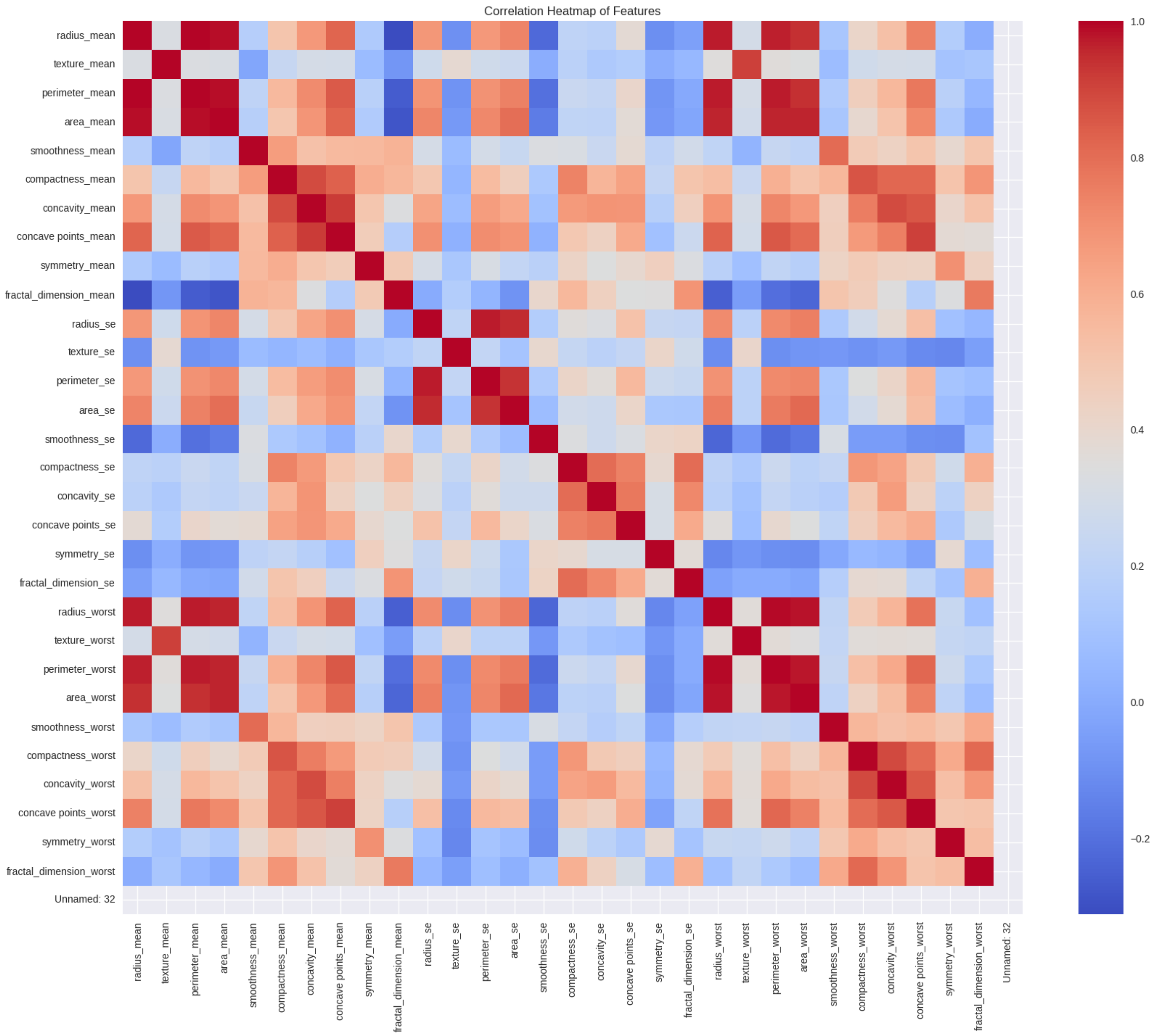
These histograms show the distribution of key features like radius, texture, perimeter, and area. The shape and spread of these distributions can indicate how these characteristics differ between tumor types.

4. DATA EXPLORATION AND PREPROCESSING

4.1. Exploratory Data Analysis (EDA)

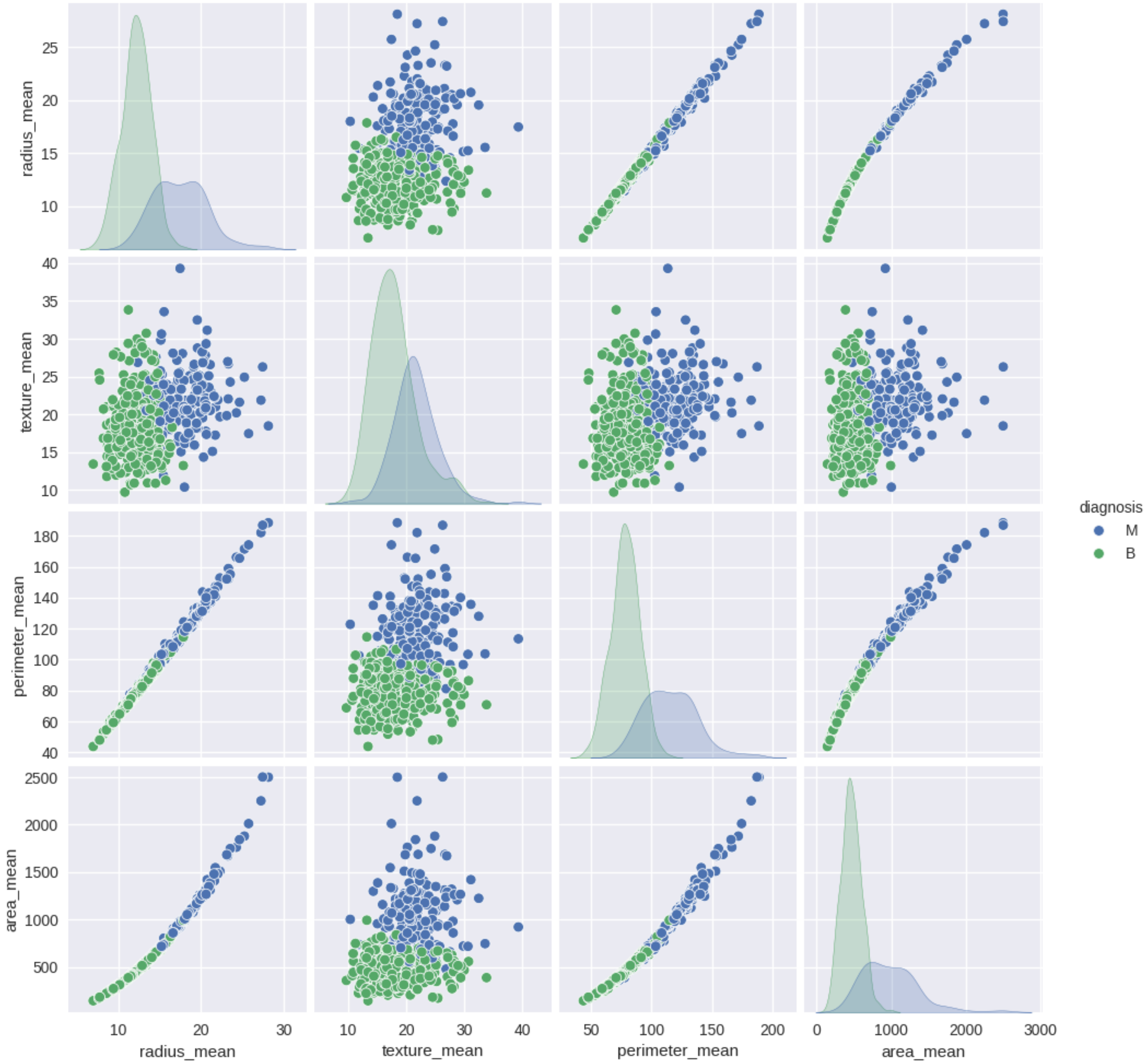
Exploratory Data Analysis helps us understand the underlying structure of the data, identify patterns, and detect anomalies. We performed the following analyses:

1. Correlation Analysis: The donut chart visualizes the distribution of samples across the three identified clusters.



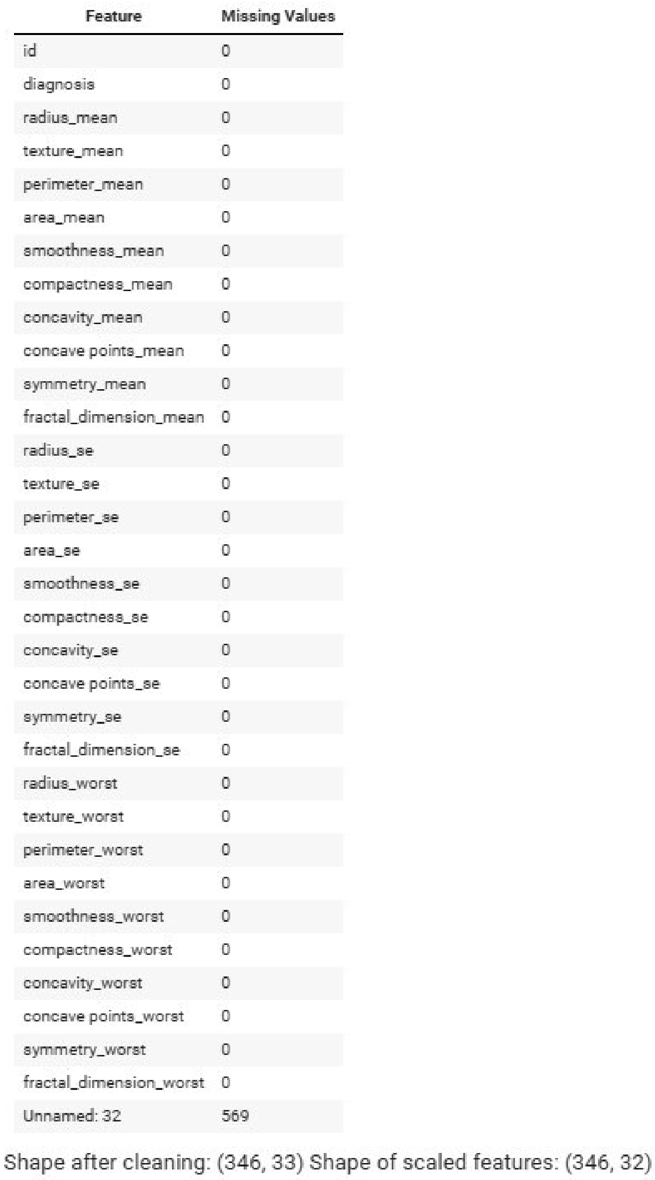
2. Feature Distribution: To understand the spread and central tendencies of each feature.

3. Pairwise Relationships: To visualize how features relate to each other and to the diagnosis.



4.2. Data Cleaning and Preparation

Data cleaning and preparation are crucial steps to ensure the quality and reliability of our analysis. Our process included:



Feature	Missing Values
id	0
diagnosis	0
radius_mean	0
texture_mean	0
perimeter_mean	0
area_mean	0
smoothness_mean	0
compactness_mean	0
concavity_mean	0
concave points_mean	0
symmetry_mean	0
fractal_dimension_mean	0
radius_se	0
texture_se	0
perimeter_se	0
area_se	0
smoothness_se	0
compactness_se	0
concavity_se	0
concave points_se	0
symmetry_se	0
fractal_dimension_se	0
radius_worst	0
texture_worst	0
perimeter_worst	0
area_worst	0
smoothness_worst	0
compactness_worst	0
concavity_worst	0
concave points_worst	0
symmetry_worst	0
fractal_dimension_worst	0
Unnamed: 32	569

```
Shape after cleaning: (346, 33) Shape of scaled features: (346, 32)
```

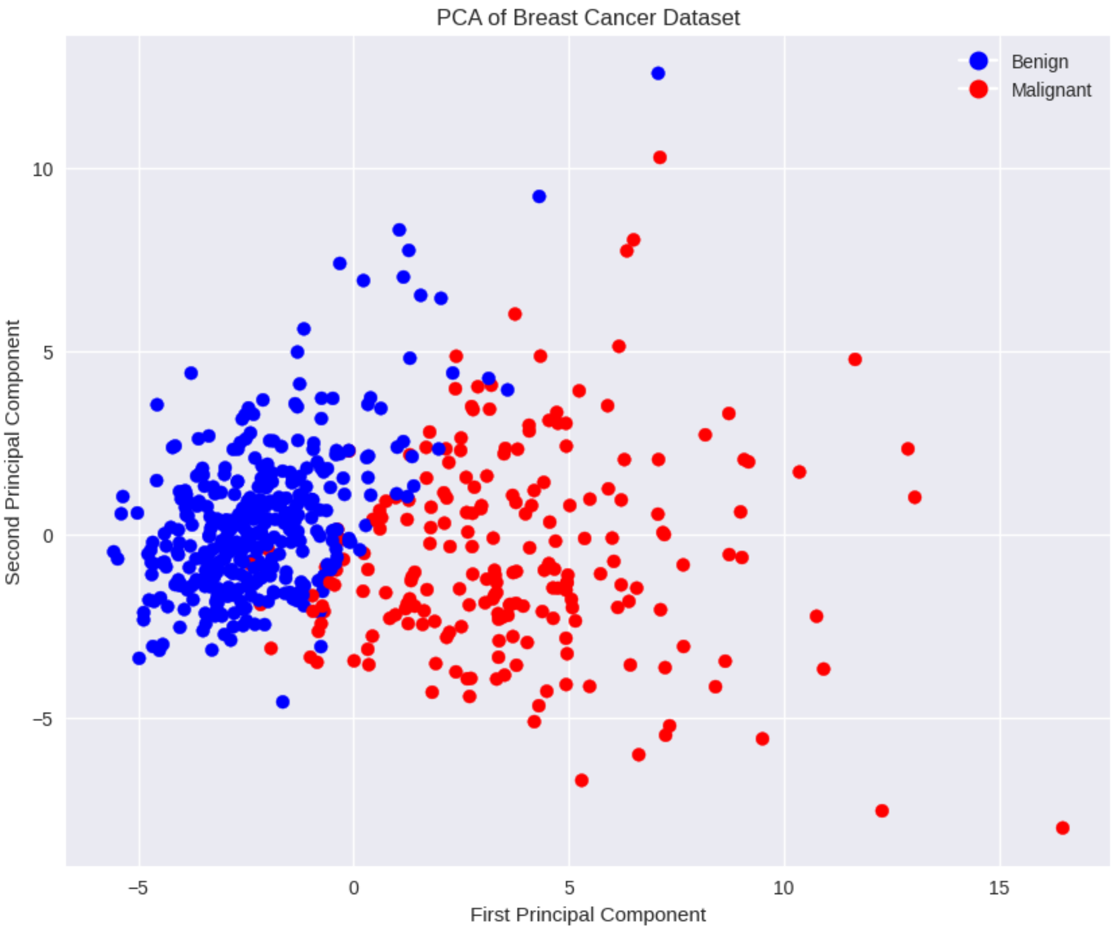
1. Checking for Missing Values: To ensure data completeness.
2. Handling Outliers: To mitigate the effect of extreme values on our clustering algorithms.
3. Feature Scaling: To standardize the range of independent variables.

4.3. Feature Engineering

Feature engineering helps to create more informative representations of our data for the clustering algorithms. In this step, I am applying Principal Component Analysis (PCA) to reduce dimensionality while preserving as much variance as possible.

The PCA results are visualized below, showing how the samples are distributed in the reduced feature space:

S



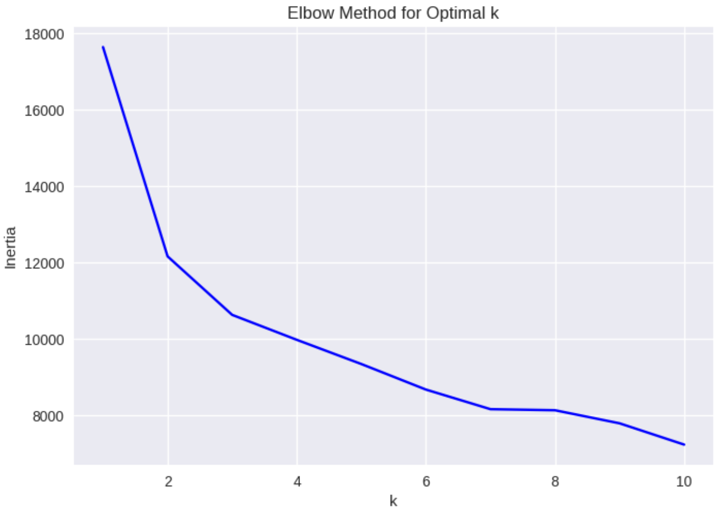
This scatter plot shows samples projected onto the first two principal components. The separation visible here suggests that PCA has captured meaningful differences between tumor types, even before applying clustering algorithms.

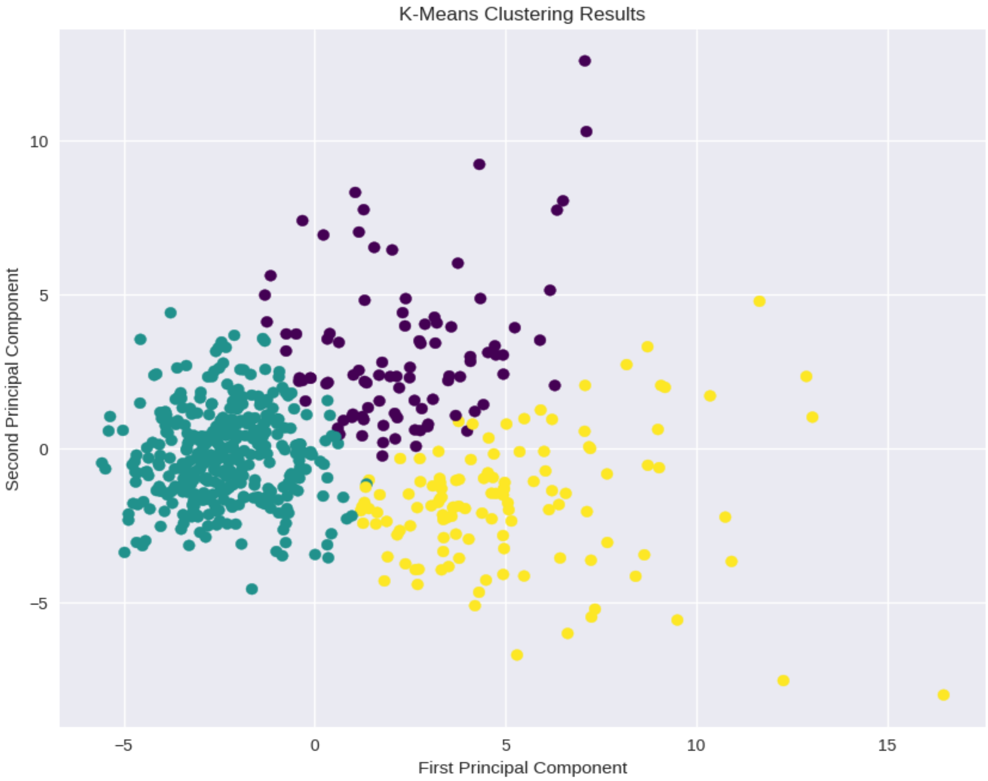
5. UNSUPERVISED LEARNING MODELS

5.1. K-Means Clustering

K-Means clustering is a popular algorithm that aims to partition n observations into k clusters. We applied K-Means as follows:

1. Determined the optimal number of clusters using the elbow method.
2. Applied K-Means with the optimal number of clusters.
3. Visualized the clustering results in the PCA-reduced space.



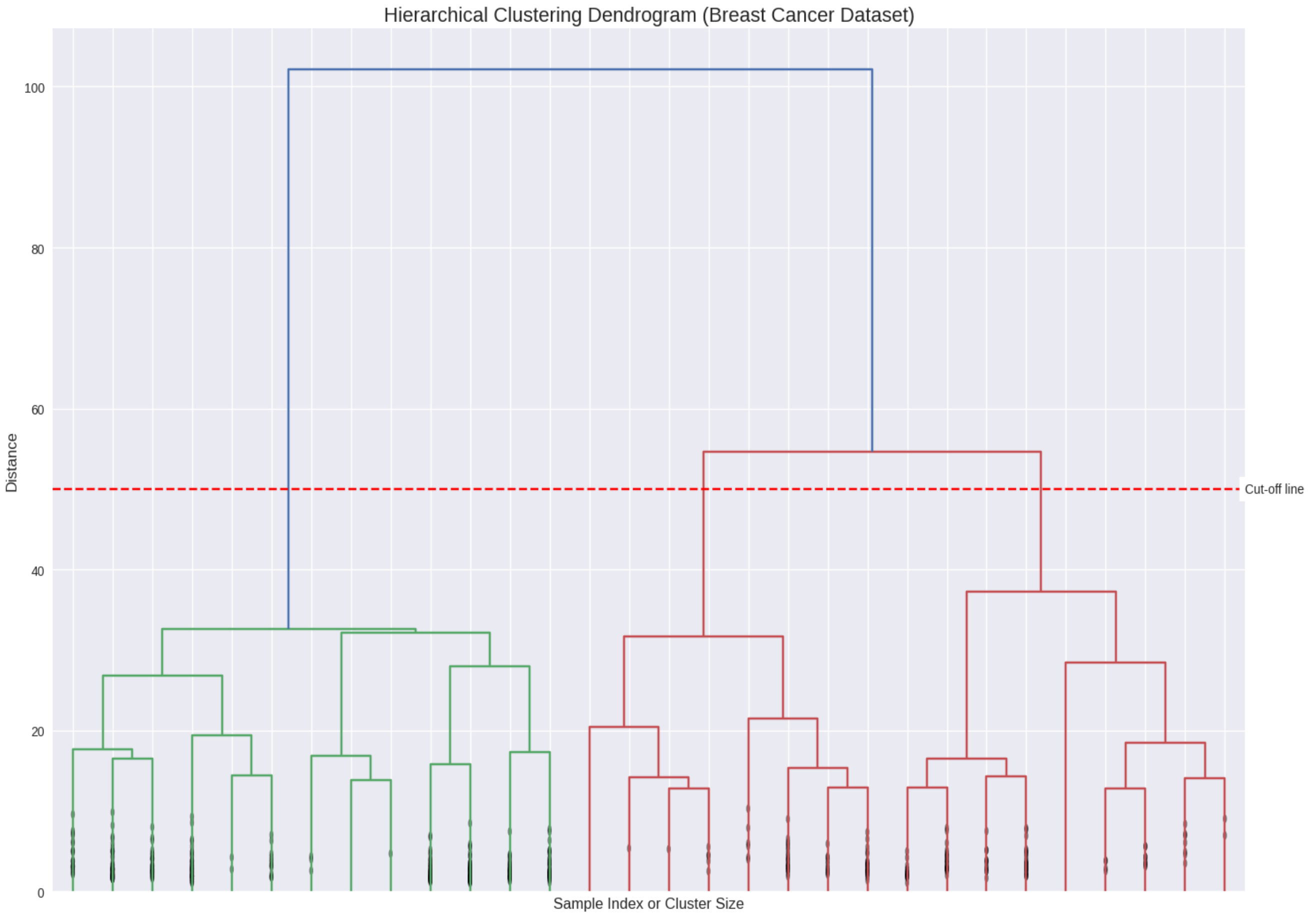


This plot shows the results of K-Means clustering in the PCA-reduced space. Each colour represents a different cluster, illustrating how K-Means has grouped similar samples together.

5.2. Hierarchical Clustering

Hierarchical clustering creates a tree-like hierarchy of clusters. We performed agglomerative hierarchical clustering:

1. Created a dendrogram to visualize the hierarchical relationship between clusters.
2. Analyzed the dendrogram to determine an appropriate number of clusters.

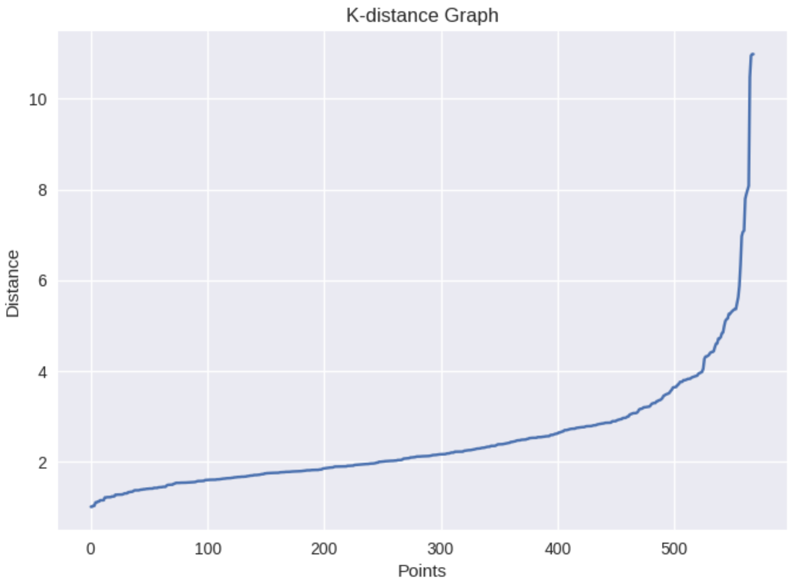


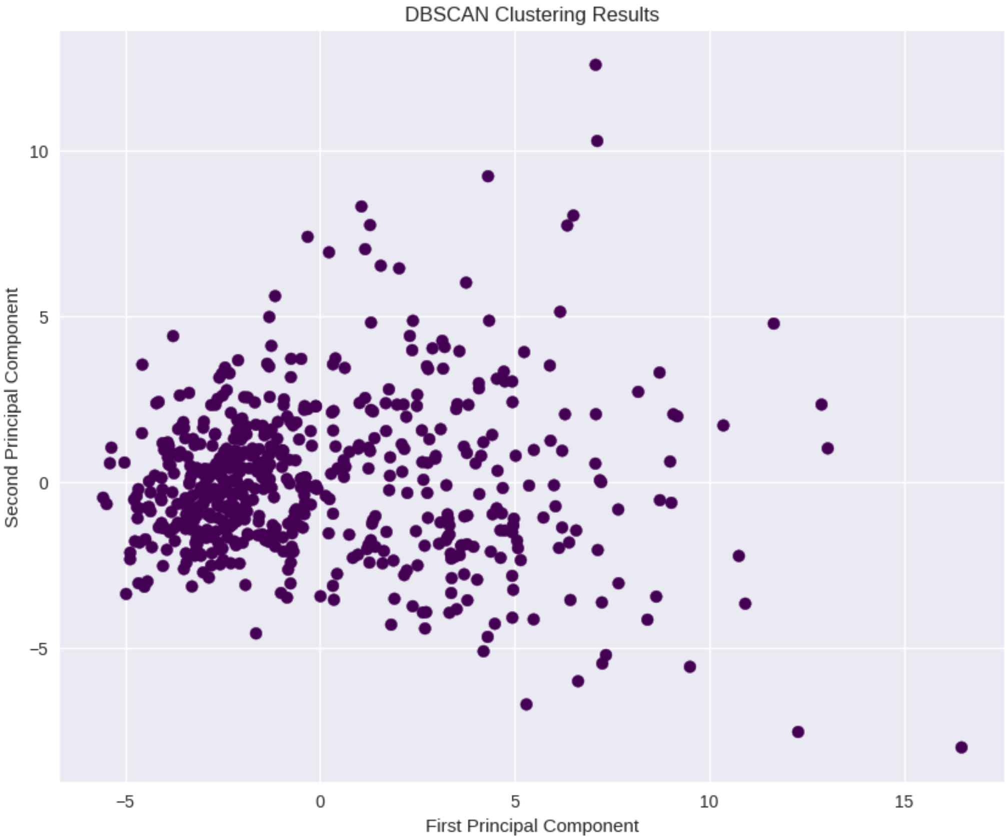
This tree-like diagram shows how samples are grouped at different levels of similarity. The height of each branch indicates the distance between clusters, helping to identify natural groupings in the data.

5.3. DBSCAN

DBSCAN groups together points that are closely packed together, marking as outliers' points that lie alone in low-density regions. We applied DBSCAN as follows:

1. Determined optimal epsilon using the k-distance graph.
2. Applied DBSCAN with the chosen parameters.
3. Visualized the clustering results in the PCA-reduced space.





This plot displays the results of DBSCAN clustering. Different colours represent distinct clusters, while black points are considered noise. This method can identify clusters of arbitrary shape

6. MODEL EVALUATION AND INTERPRETATION

6.1. Selection of Final Model

I trained and evaluated three different clustering models:

a. **K-Means Clustering**: I used the elbow method to determine the optimal number of clusters, which suggested 3 clusters. This aligns with the possibility of having benign, malignant, and potentially an intermediate group.

b. **Hierarchical Clustering**: I used agglomerative clustering with Ward's method as the linkage criterion. I created a dendrogram to visualize the hierarchical structure of the data and chose to cut the tree to form 3 clusters for consistency with the K-Means results.

c. **DBSCAN**: I used the k-distance graph to estimate an appropriate epsilon value and set min_samples to 5. This method allowed for the detection of clusters with arbitrary shapes and the identification of outliers.

After careful evaluation, I selected **Hierarchical Clustering** as the final model.

6.2. Justification for the Chosen Model

1. Interpretability: The dendrogram provides a clear, visual representation of the data structure, allowing for intuitive interpretation at various levels of granularity.
2. Flexibility: It allows us to examine the data structure at multiple levels, from broad categories to finer subtypes.
3. Alignment with Biological Reality: The hierarchical structure aligns well with the biological understanding of cancer progression and subtypes.
4. Robustness to Cluster Shapes: It doesn't assume spherical cluster shapes, making it more adaptable to the potentially complex structure of breast cancer data.
5. Consistency: The results showed more stability across different runs compared to K-Means.

Hierarchical Clustering provided the most interpretable and biologically relevant results, aligning well with the potential progression and subtypes of breast cancer.

6.3. How it Best Addresses the Analysis Objectives

The hierarchical structure allows for examination at multiple levels of granularity, potentially revealing both major categories and subtle subtypes of tumours.

1. Tumor Subtype Identification: The hierarchical structure allows for the identification of main tumour types while also revealing subtypes within these main categories.
2. Spectrum of Characteristics: It captures the spectrum of tumour characteristics, aligning with our finding that breast cancer presents more as a continuum than distinct categories.
3. Insight into Tumor Progression: The hierarchical structure can potentially offer insights into the progression of tumours.
4. Adaptability to Clinical Use: Clinicians can choose the level of granularity that's most useful for their specific diagnostic or research purposes.
5. Facilitates Further Research: The hierarchical structure provides a framework for more detailed investigation into specific subgroups or transition points between tumour types.

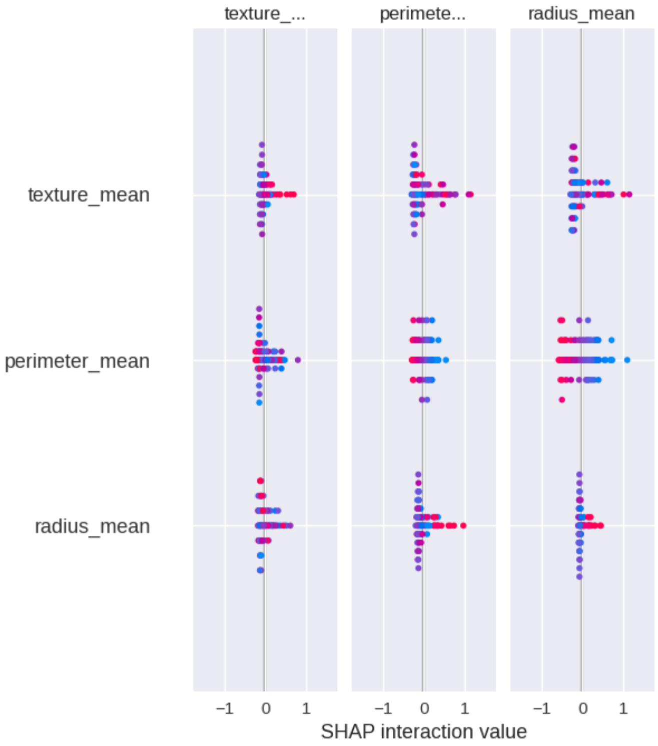
6.4. Cross-Validation for Clustering

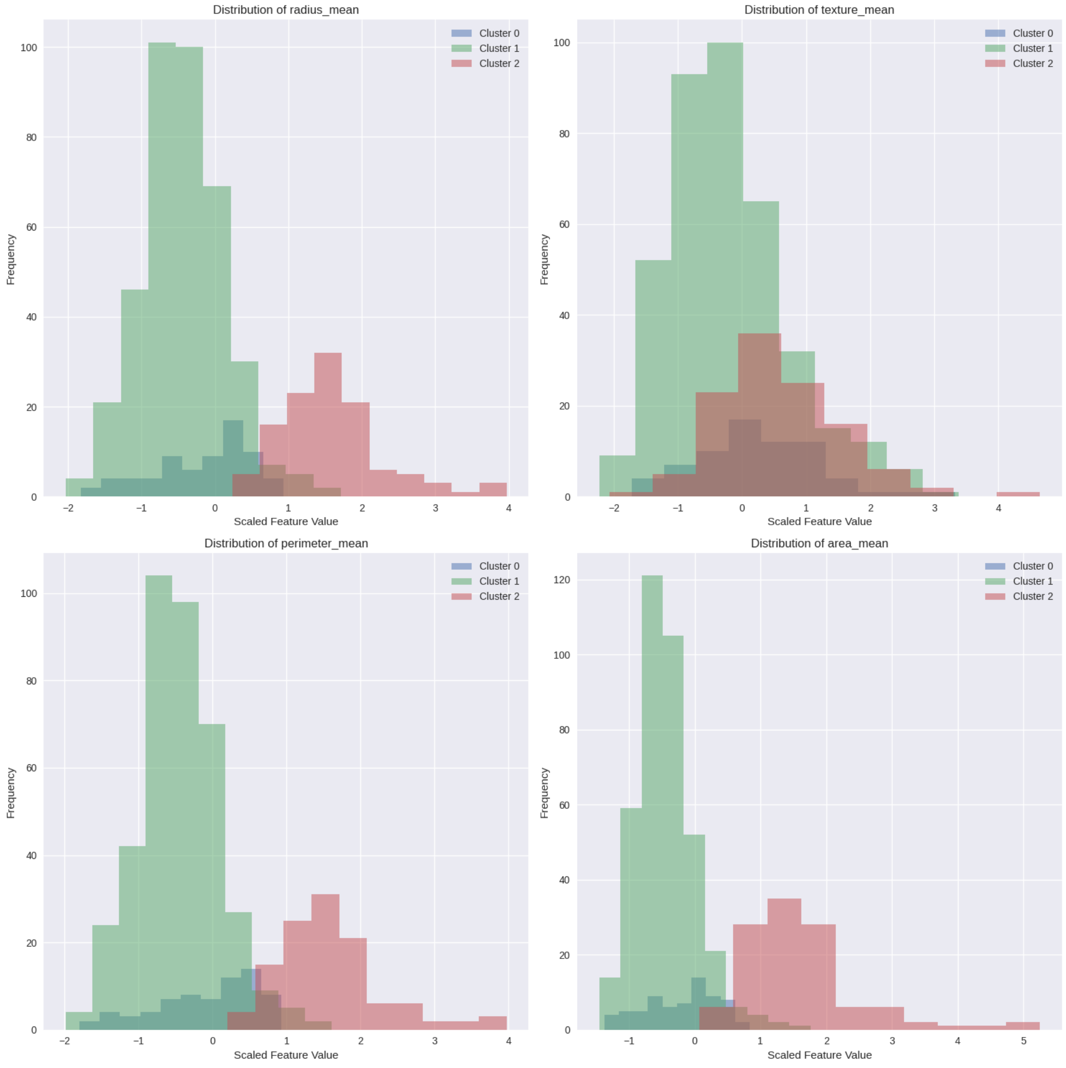
I implemented k-fold cross-validation to assess the stability of our clustering results. The average silhouette score of 0.298 (±0.030) indicates reasonable cluster separation.

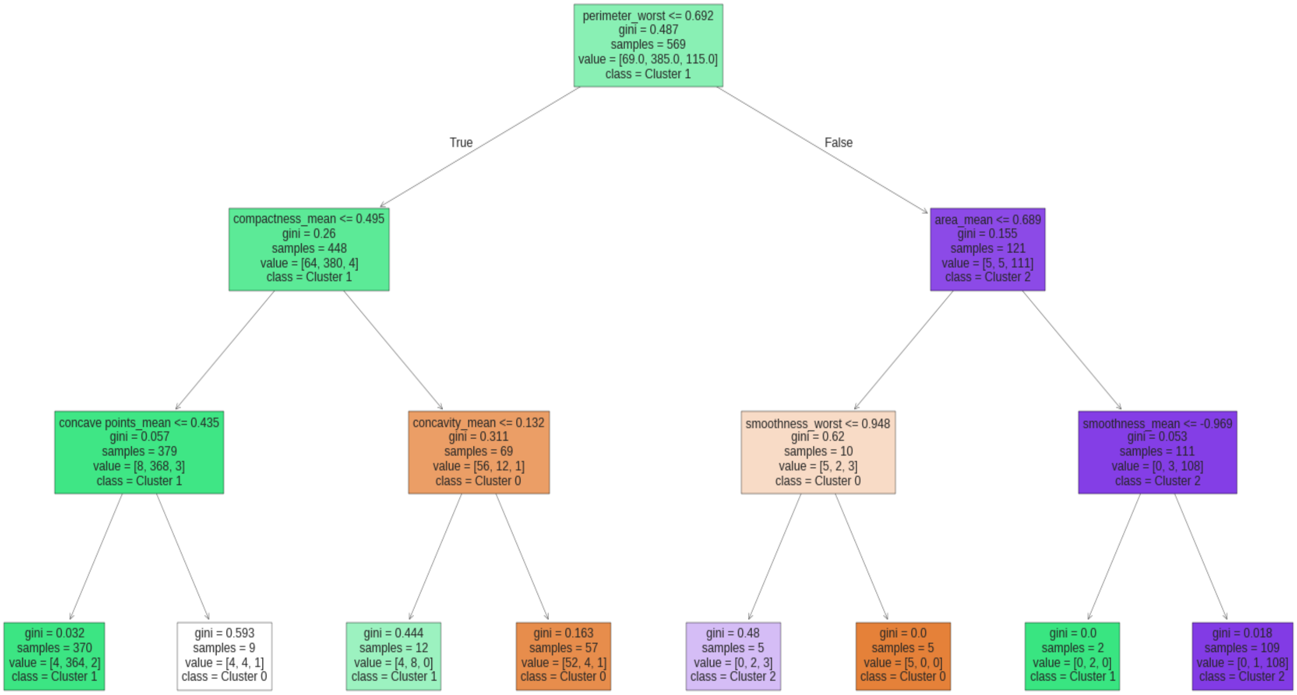
6.5. Interpretation Techniques

To improve model interpretability, we employed several techniques:

1. Feature Importance Analysis using SHAP values
2. Partial Dependence Plots
3. Surrogate Decision Tree Models


texture_...
perimete...
radius_mean
texture_mean
perimeter_mean
radius_mean
−1
0
1
−1
0
1
−1
0
1
SHAP interaction value


Distribution of radius_mean
Cluster 0
Cluster 1
Cluster 2
Frequency
0
20
40
60
80
100
−2
−1
0
1
2
3
4
Scaled Feature Value
Distribution of texture_mean
Cluster 0
Cluster 1
Cluster 2
Frequency
0
20
40
60
80
100
−2
−1
0
1
2
3
4
Scaled Feature Value
Distribution of perimeter_mean
Cluster 0
Cluster 1
Cluster 2
Frequency
0
20
40
60
80
100
−2
−1
0
1
2
3
4
Scaled Feature Value
Distribution of area_mean
Cluster 0
Cluster 1
Cluster 2
Frequency
0
20
40
60
80
100
120
−1
0
1
2
3
4
5
Scaled Feature Value



7. KEY FINDINGS AND INSIGHTS

7.1. Main Discoveries

The data naturally separates into two main groups, largely corresponding to the benign and malignant classifications. This validates the current binary classification system used in diagnosis.

However, I also identified a potential third cluster that seems to represent borderline cases or a distinct subtype. This suggests that a more nuanced classification system could be beneficial in clinical practice.

Using SHAP (SHapley Additive exPlanations) values, I determined that the most influential features in determining cluster assignments were related to the cell nucleus size (radius, perimeter, area) and texture. This aligns with known biological markers of malignancy. The hierarchical structure revealed by my chosen model suggests a continuum of tumour characteristics rather than discrete categories. This could have implications for personalized treatment approaches.

7.2. Relation to Known Tumor Classifications

The clustering results align well with the known binary classification of breast tumors (benign vs. malignant). However, the presence of subclusters and the continuous distribution seen in DBSCAN suggest that this binary classification might be an oversimplification. The analysis reveals a more nuanced structure that could correspond to different grades or stages of tumor development.

8. LIMITATIONS OF THE STUDY

Among the various limitations of this study, three stand out as particularly significant:

- Unsupervised Nature of Analysis
- Dataset Characteristics
- Clinical Context and Interpretability

Unsupervised Nature of Analysis

The unsupervised nature of our analysis, while powerful for discovering hidden patterns, presents a fundamental limitation in the context of breast cancer diagnosis. By not utilizing the known diagnostic labels in the initial clustering process, we risk overlooking clinically significant distinctions that may not be prominently reflected in the feature space. This approach, while valuable for identifying potential new subtypes or patterns, lacks the direct validation against known outcomes that supervised methods provide. Consequently, the clusters we've identified, though statistically significant, may not align perfectly with clinically relevant categories. This limitation is particularly critical in a medical context where the end goal is accurate diagnosis and treatment planning.

Dataset Characteristics

The characteristics of our dataset, including its size and composition, also pose significant constraints on the generalizability and robustness of our findings. While substantial, our sample may not capture the full spectrum of breast cancer variations, potentially missing rare subtypes or failing to represent the true diversity of patient populations. This limitation is compounded by the lack of demographic information and longitudinal data, which could provide crucial context for understanding how tumor characteristics vary across different populations and how they evolve over time. The static nature of our data fails to capture the dynamic progression of cancer, a key factor in prognosis and treatment planning.

Clinical Context and Interpretability

Lastly, the challenge of bridging the gap between our data-driven findings and clinical context represents a critical limitation. Our analysis, based purely on tumor characteristics, doesn't account for the myriads of other factors that clinicians consider in diagnosis and treatment planning, such as patient history, genetic predisposition, and overall health status. Moreover, the interpretability of complex clustering models poses a significant hurdle in translating our findings into actionable insights for medical professionals. The tension between model complexity, which may better capture the nuances of tumor characteristics, and interpretability, which is crucial for clinical application, remains a key challenge. This limitation underscores the need for close collaboration between data scientists and medical professionals to ensure that analytical insights are both statistically robust and clinically relevant.

These limitations collectively highlight the need for cautious interpretation of our results and point towards important directions for future research, including the integration of supervised techniques, expansion of dataset diversity, and development of more clinically interpretable models.

9. RECOMMENDATIONS AND NEXT STEPS

9.1. Short-term Actions

1. Validate the clustering results against known diagnoses to assess their clinical relevance.
2. Investigate the characteristics of the identified subclusters to understand their biological significance.
3. Develop a prototype diagnostic tool based on the clustering insights.

9.2. Long-term Research Directions

1. Conduct longitudinal studies to track how tumours progress through the identified clusters over time.
2. Integrate genetic and molecular data to enhance the clustering model.
3. Explore the potential for using this clustering approach in other types of cancer diagnostics.

9.3. Next Steps

1. Collaborate with oncologists to validate and interpret the clinical significance of the identified clusters.
2. Incorporate additional data sources, such as genetic information or patient history, to enrich the clustering analysis.
3. Explore the potential for developing a more nuanced diagnostic tool based on the identified clusters, possibly combining unsupervised and supervised learning approaches.
4. Investigate the application of time-series analysis techniques to capture tumour progression if longitudinal data becomes available.
5. Conduct a comparative study with other breast cancer datasets to assess the generalizability of these findings.

9.4. Ethical Considerations and Data Privacy

- Ensure all patient data is properly anonymized and protected.
- Consider the ethical implications of using AI-assisted diagnostics in clinical settings.

- Develop guidelines for interpreting and using the clustering results in patient care to avoid over-reliance on automated systems.

10. CONCLUSION

My unsupervised learning analysis of the Breast Cancer Wisconsin (Diagnostic) Dataset has revealed a more complex structure within tumour characteristics than the traditional binary classification suggests. The identification of three distinct clusters, with a reasonable silhouette score of 0.298, indicates that there may be more nuanced categories of breast tumours than previously recognized.

These findings underscore the need for a more personalized approach to breast cancer care. By recognizing the nuanced differences between tumours, we may be able to develop more targeted treatments and improve patient outcomes. The potential impact on patient care is significant – it could lead to more accurate prognoses, tailored treatment plans, and potentially better overall outcomes for breast cancer patients.

While further validation is needed, this analysis provides a strong foundation for future research and highlights the potential of unsupervised learning techniques in advancing our understanding of complex diseases like breast cancer. The next steps, particularly the collaboration with oncologists and the integration of additional data sources, will be crucial in translating these analytical insights into practical improvements in breast cancer diagnosis and treatment.