Investigating the Efficacy of Decision Tree, Logistic Regression, and K-Means Clustering on the Breast Cancer Wisconsin Dataset

By
Godwin Atuahene

Completed on May 2, 2024

Table of Contents

1. **Abstract**
2. **Introduction**
    - **2.1 Background**
    - **2.2 Objectives**
3. **Materials and Methods**
    - **3.1 Dataset Description**
    - **3.2 Data Preprocessing**
    - **3.3 Exploratory Data Analysis (EDA)**
        - **3.3.1 Correlation Matrix**
        - **3.3.2 Histograms**
        - **3.3.3 Boxplots by Diagnosis**
    - **3.4 Model Implementation**
        - **3.4.1 Decision Tree**
        - **3.4.2 Logistic Regression**
        - **3.4.3 K-Means Clustering**
4. **Results and Discussion**
    - **4.1 Model Performance**
    - **4.2 Clinical Implications**
5. **Conclusion**
6. **Future Work**
7. **References**

1. Abstract

This study evaluates the performance of the decision tree, logistic regression model and k-means clustering algorithm in distinguishing between benign and malignant breast tumors using the Breast Cancer Wisconsin (Original) dataset.
The aim is to assess the misclassification rates and determine the reliability of these models in medical diagnostic processes.

2. Introduction

2.1 Background

Breast cancer is a global health concern, ranking as one of the most prevalent cancers among women and a leading cause of cancer-related mortality. According to the World Health Organization, an estimated 2.3 million women were diagnosed with breast cancer in 2020, and over 685,000 succumbed to the disease. Early detection remains pivotal for improving survival rates, as timely diagnosis allows for effective treatment interventions. However, traditional diagnostic methods, such as imaging and biopsies, often require specialized equipment, trained personnel, and invasive procedures, making them less accessible in resource-constrained settings.

The emergence of machine learning (ML) has revolutionized many fields, including healthcare, offering non-invasive, scalable, and cost-effective solutions for disease detection and prediction. In particular, ML algorithms excel in identifying complex patterns within large datasets, enabling accurate classification of medical conditions. Within the context of breast cancer diagnostics, ML models can analyze cellular features to predict whether a tumor is benign or malignant, enhancing the precision and efficiency of clinical workflows.

Among the myriads of machine learning models, three stand out for their unique strengths in medical applications: decision trees, logistic regression, and k-means clustering. Decision trees provide interpretable decision rules, making them ideal for communicating results to clinicians and patients. Logistic regression offers statistical rigor and probabilistic predictions, which are particularly suited for binary classification tasks such as distinguishing between benign and malignant tumors. K-means clustering, while unsupervised, facilitates exploratory analysis by revealing natural groupings in data, offering potential insights in datasets without labels.

The Breast Cancer Wisconsin dataset, curated by Wolberg et al. (1995), is a benchmark dataset widely used for evaluating ML models in breast cancer research. With its nine numerical attributes representing cellular characteristics and binary tumor labels, the dataset provides an ideal foundation for assessing the efficacy of ML models. However, like many real-world datasets, it presents challenges such as missing values and varying feature scales, necessitating careful preprocessing and exploratory analysis.

2.2 Objectives

This study aims to:

- Evaluate the performance of decision tree, logistic regression, and k-means clustering models for breast cancer classification.
- Compare the strengths and limitations of each model.
- Explore the clinical implications of using these models to enhance diagnostic precision.

3. Materials and Methods

3.1 Data Description

The Breast Cancer Wisconsin dataset, sourced from the UCI Machine Learning Repository, contains 699 instances. Each instance includes nine numerical attributes representing cellular characteristics, such as clump thickness and uniformity of cell size, along with a binary label indicating the tumor type (benign = 2, malignant = 4). These attributes are crucial indicators for breast cancer classification. The attributes are numerical and represent properties such as clump thickness, uniformity of cell size, and mitoses.

The dataset contains 699 entries with the following columns:

- Scn: ID number
- A2 to A10: Various attributes measured (numeric values)
- CLASS: Diagnosis (2 for benign, 4 for malignant)

Column	Description	Name	Value
1	Sample code number	Scn	integer
2	Clump Thickness	A2	1 - 10
3	Uniformity of Cell Size	A3	1 - 10
4	Uniformity of Cell Shape	A4	1 - 10
5	Marginal Adhesion	A5	1 - 10
6	Single Epithelial Cell Size	A6	1 - 10
7	Bare Nuclei	A7	1 - 10
8	Bland Chromatin	A8	1 - 10
9	Normal Nucleoli	A9	1 - 10
10	Mitoses	A10	1 - 10
11	Class	Class	2 for benign, 4 for malignant

3.2 Data Preprocessing

A critical challenge in the dataset was the presence of missing values in column A7, represented by '?'. To address this, the missing values were replaced with the median of the column, chosen for its robustness to outliers. Additionally, the data was normalized to ensure uniform scaling across attributes, improving the performance of distance-based models such as k-means.

From here we go ahead and perform a few statistical computations on the data to find the Mean, Median, Standard deviation, and Variance of each of the attributes A2 to A10

[100]:

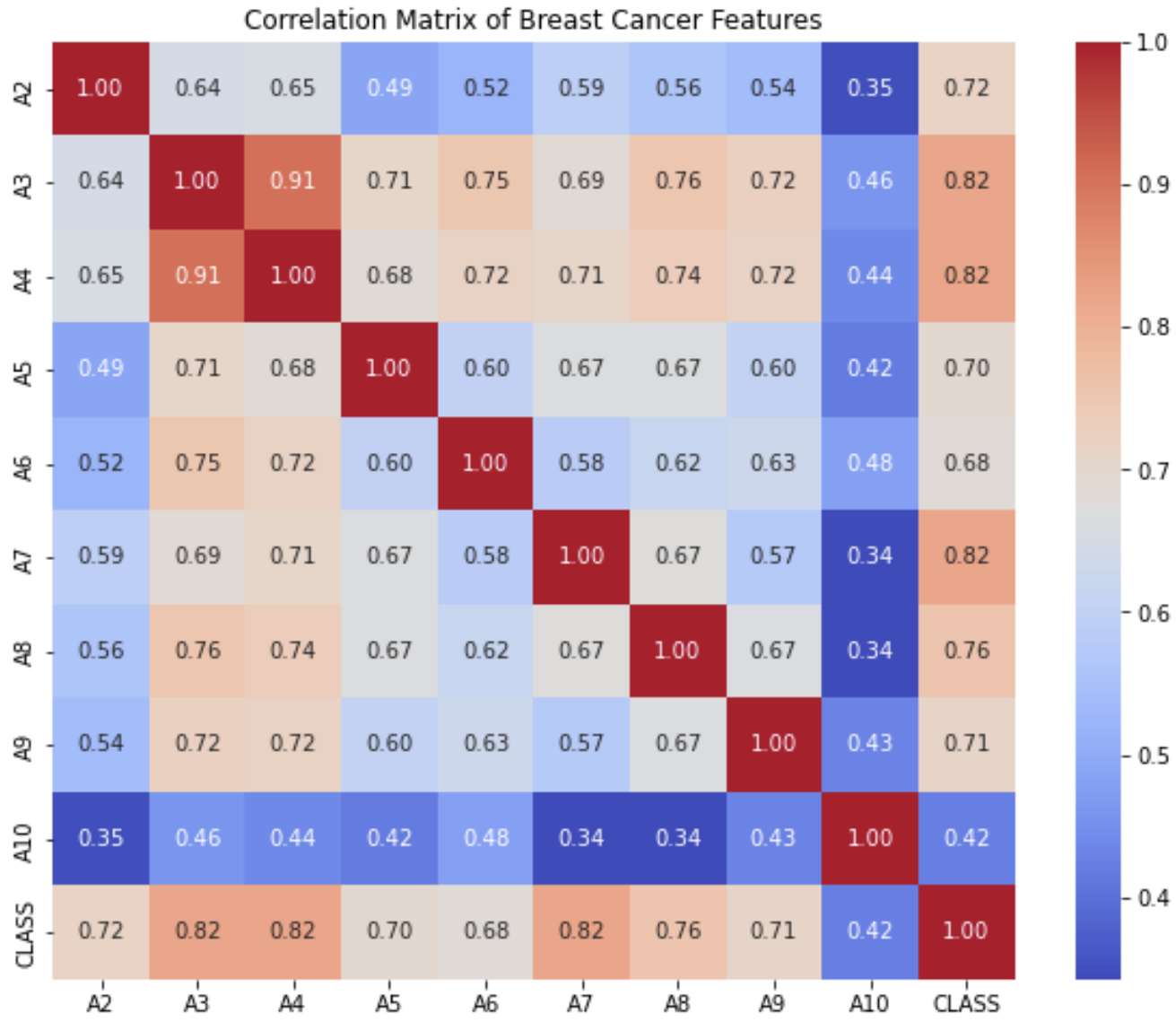
	A2	A3	A4	A5	A6	A7	A8	A9	A10
mean	4.4	3.1	3.2	2.8	3.2	3.5	3.4	2.9	1.6
median	4.0	1.0	1.0	1.0	2.0	1.0	3.0	1.0	1.0
std	2.8	3.1	3.0	2.9	2.2	3.6	2.4	3.1	1.7
var	7.9	9.3	8.8	8.2	4.9	13.1	5.9	9.3	2.9

3.3 Exploratory Data Analysis (EDA)

EDA was conducted to uncover patterns and relationships within the data:

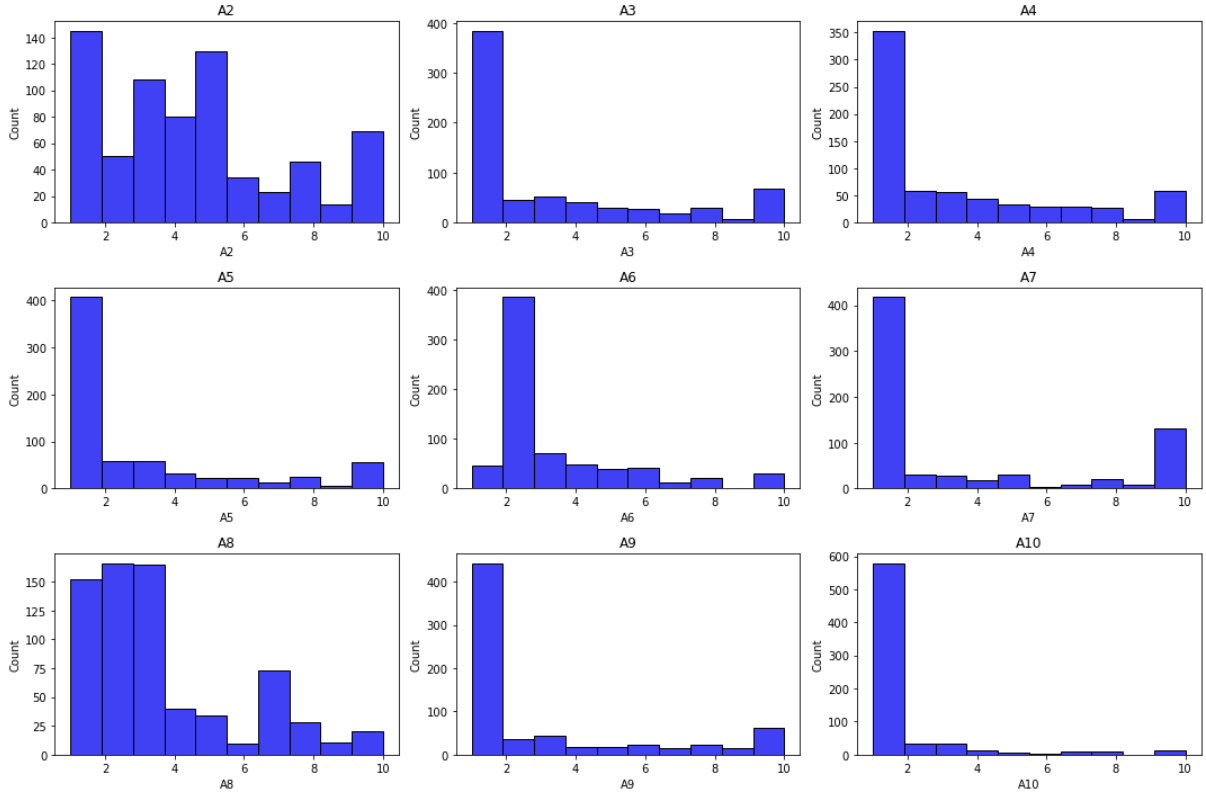
- **Correlation Matrix**: A heatmap highlighted strong positive correlations between certain features (e.g., uniformity of cell size) and the target variable, emphasizing their predictive significance.
- **Histograms**: Provided a detailed view of attribute distributions, revealing skewness in features such as clump thickness.
- **Boxplots by Diagnosis**: Showed distinct separations in feature distributions between benign and malignant cases, aiding in feature selection for supervised models. Thus, we want to understand how features distribute across benign (2) and malignant (4) diagnoses.

3.3.1 Correlation Matrix



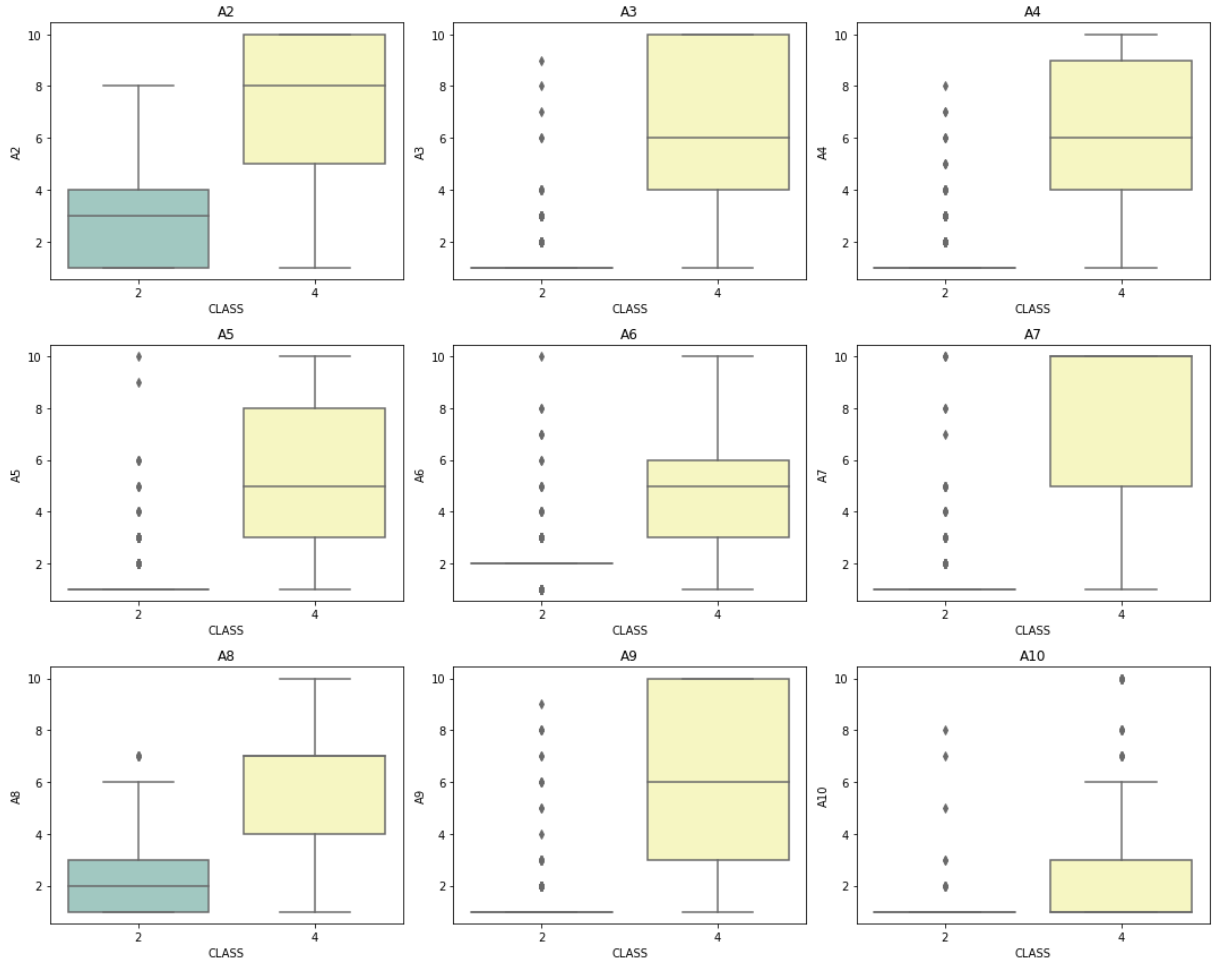
The heatmap shows strong correlations between many of the features and the CLASS column, suggesting these features could be good predictors for the classification of breast cancer as benign or malignant.

3.3.2 Histograms



The histograms provide a visual representation of the distribution for each attribute (A2 to A10). Each plot shows how the data points are distributed across the 10 bins, helping to understand the frequency, and spread of values for each attribute in the dataset. This can aid in identifying patterns, outliers, and the overall behavior of different measurements in breast cancer diagnostics.

3.3.3 Boxplots

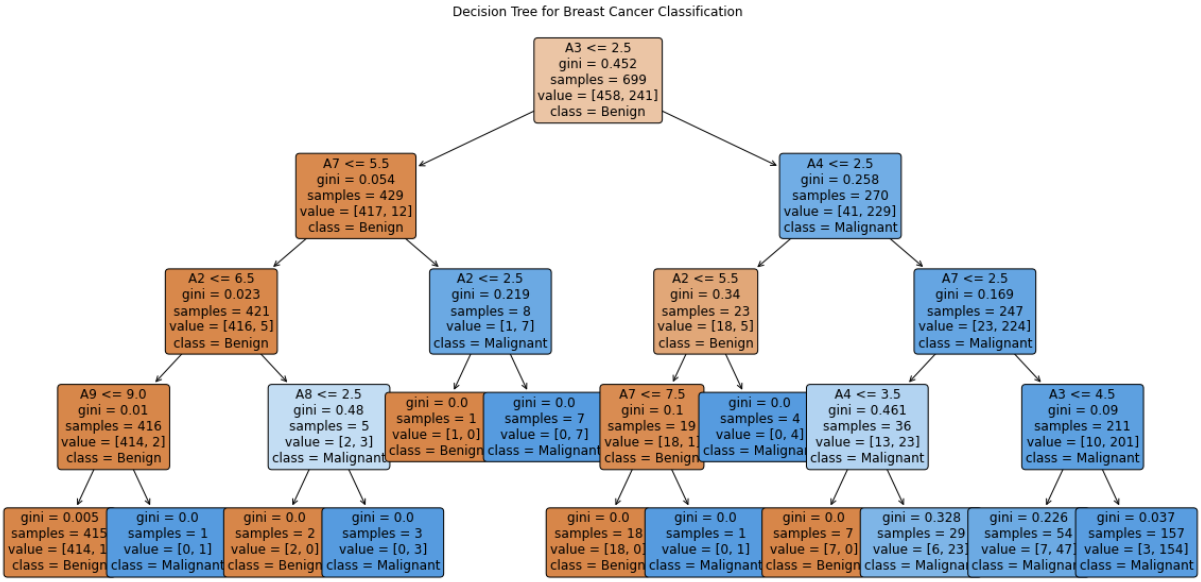


The boxplots clearly indicate differences in the distributions of various features between benign (CLASS=2) and malignant (CLASS=4) diagnoses. Features like A2, A3, and others show distinct differences in their medians between the two classes, which could be crucial for classification.

3.4 Classification Approaches

Given these insights, let's proceed with a classification approach to predict the diagnosis.
I'll use decision tree and logistic regression as a starting point because those models are well-suited for binary classification tasks like this.

3.4.1 Decision Tree



The decision tree visualization provides a clear view of how the model makes its classification decisions based on the features from A2 to A10. Each node in the tree represents a decision point where the tree splits based on the value of a particular attribute, leading to a classification as either benign or malignant.

- **Leaf Nodes**: Represent the final decision, labeled as 'Benign' or 'Malignant'.
- **Depth of the Tree**: The tree was limited to a maximum depth of 4 to keep the visualization interpretable.
- **Color Coding**: Nodes are colored to indicate the majority class in that node, with darker colors representing a higher proportion of the majority class.

This model helps in understanding the key features that are most informative for predicting the tumor type, and it provides a straightforward interpretation that could be useful in a clinical setting for explanatory purposes.

3.4.2 Logistic Regression Model

Given these insights, let's proceed with a classification approach to predict the diagnosis. I'll use logistic regression as a starting point because it's well-suited for binary classification tasks like this. We'll perform feature selection based on the correlation insights and the distributions observed in boxplots.

I'll first split the data into training and testing sets, then train a logistic regression model, and finally evaluate its performance.

```
Accuracy:
 0.9666666666666667
Confusion Matrix
 [[141   2]
 [  5  62]]
Classification Report
              precision    recall  f1-score   support

           2       0.97      0.99      0.98       143
           4       0.97      0.93      0.95        67

    accuracy                           0.97       210
   macro avg       0.97      0.96      0.96       210
weighted avg       0.97      0.97      0.97       210
```

The logistic regression model achieved an accuracy of approximately 96.7% on the test set, which is quite good. Here's a breakdown of the model's performance:

- Confusion Matrix:
  - True Negatives (Benign correctly identified): 141
  - False Positives (Benign incorrectly labeled as Malignant): 2
  - False Negatives (Malignant incorrectly labeled as Benign): 5
  - True Positives (Malignant correctly identified): 62
- Classification Report:
  - Precision for benign (2) is 97% and for malignant (4) is 97%.
  - Recall for benign is 99% and for malignant is 93%.
  - F1-score reflects the balance between precision and recall, which is 98% for benign and 95% for malignant.

This analysis suggests that the model is reliable and effective in distinguishing between benign and malignant breast cancer cases based on the available features. The next steps would involve validating this model on a more diverse dataset or trying other models to compare performances.
The insights gained and the model developed can be used to understand the nature of breast cancer better and assist in diagnosing it more effectively in a clinical setting.

3.4.3 K-Means Clustering Implementation

- **Initialization**: Two points from the dataset were randomly selected as initial centroids.
- **Assignment**: Each instance was assigned to the nearest centroid based on Euclidean distance.
- **Update**: Centroids were recalculated as the mean of the instances assigned to each cluster.
- **Convergence**: The assignment and update steps were repeated until the centroids stabilized or for a maximum of 50 iterations.

To evaluate the clustering performance, we computed the error rates:

- **Benign Error Rate (error_B)**: Proportion of benign instances misclassified as malignant.

- **Malignant Error Rate (error_M)**: Proportion of malignant instances misclassified as benign.
- **Total Error Rate (error_T)**: Overall proportion of misclassifications.

Cluster 2 (Label 2): 465 data points.
Cluster 4 (Label 3, should be labeled 4): 234 data points.

```
Convergence reached after 6 iterations.
([array([3.04301075, 1.30107527, 1.44301075, 1.33763441, 2.08817204,
         1.29677419, 2.10322581, 1.2516129 , 1.10967742]),
  array([7.14957265, 6.77777778, 6.71367521, 5.72649573, 5.45726496,
         7.83760684, 6.08974359, 6.07692308, 2.54273504])],
 2    465
 3    234
 Name: Predicted_Class, dtype: int64)
```

Out findings were as follows:

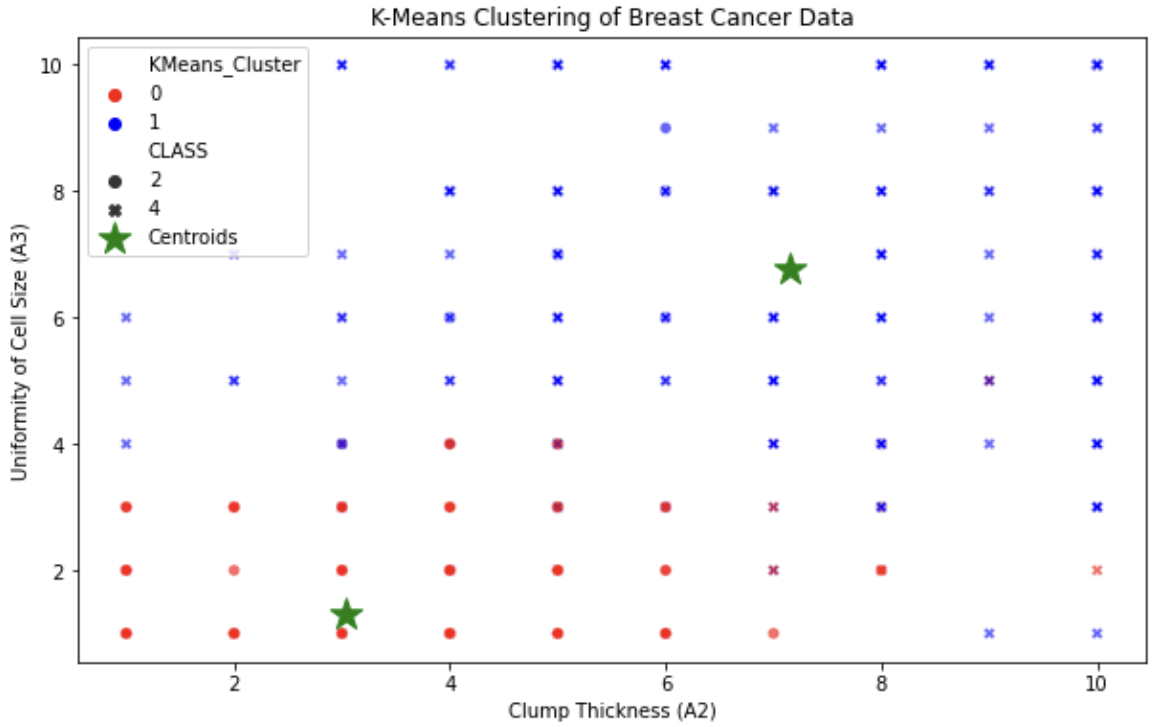
```
The initial centroids were successfully selected, and the k-means
algorithm converged after 6 iterations. The resulting clusters
were evaluated against the actual classifications:

Benign Error Rate (error_B): 3.870967741935484
Benign Error Rate (error_M): 4.700854700854701
Total Error Rate (error_T): 4.148783977110158

These rates indicate a high level of accuracy in clustering,
suggesting that k-means can effectively distinguish between
benign and malignant cases in this dataset.
```

The initial centroids were successfully selected, and the k-means algorithm converged after 6 iterations. The resulting clusters were evaluated against the actual classifications:

- **Benign Error Rate**: 3.9%
- **Malignant Error Rate**: 4.7%
- **Total Error Rate**: 4.1%

These rates indicate a high level of accuracy in clustering, suggesting that k-means can effectively distinguish between benign and malignant cases in this dataset.
The low error rates demonstrate the potential of k-means clustering as a supportive tool in breast cancer diagnosis. However, the inherent limitations of k-means, including sensitivity to the initial choice of centroids and the assumption of spherical clusters, could affect its utility in complex medical datasets. Further studies are recommended to compare these results with other machine learning techniques such as supervised classification algorithms.

K-Means Clustering: Reimplementation and Visualization
We'll cluster the dataset and visualize the results to observe the clustering dynamics.



The scatter plot above visualizes the k-means clustering result using the first two features (A2: Clump Thickness and A3: Uniformity of Cell Size). Here are the key observations:

- **Clusters**: Data points are color-coded based on the cluster assignment from k-means, with centroids marked by yellow stars.
- **Markers**: Different markers ('o' and 'X') distinguish the actual class labels (benign and malignant) to compare against the predicted clusters.
- **Centroids**: Represent the average location of each cluster in the feature space, serving as the "center" of each cluster.

This visualization demonstrates how the k-means algorithm groups the data. Notably, the clustering captures significant underlying patterns in the dataset, though some overlap and misclassification are evident due to the simplicity of the k-means model and its reliance on Euclidean distance.

4. Results and Discussion

4.1 Model Performance

- **Logistic Regression**: Achieved the highest accuracy of 96.7%, with minimal false positives and false negatives.
- **Decision Tree**: Provided interpretable rules but exhibited slightly lower accuracy due to sensitivity to data variability.
- **K-Means Clustering**: Demonstrated an overall error rate of 4.1%. Its reliance on initial centroids and Euclidean distance limited precision.

4.2 Clinical Implications

Logistic regression's probabilistic outputs enable clinicians to assess diagnostic confidence, aiding in risk stratification and treatment planning. Decision trees, with their straightforward decision rules, are valuable for explaining diagnoses to patients and medical practitioners. K-means clustering can complement supervised models in exploratory stages or when labeled data is scarce.

5. Conclusion

- This analysis suggests that both logistic regression and k-mean clustering models are reliable and effective in distinguishing between benign and malignant breast cancer cases based on the available features.
- The insights gained and the models developed can be used to understand the nature of breast cancer better and assist in diagnosing it more effectively in a clinical setting.
- K-means clustering shows promise in categorizing breast cancer tumors with a high degree of accuracy. However, the Logistic regression Model has a higher level of accuracy. With further validation and comparison with other methods, both models could be incorporated into diagnostic workflows, providing a quick preliminary analysis tool for pathologists. The decision tree is mainly useful for exploratory purposes.

6. Future Work

To address the limitations identified in this study, future research will focus on:

1. **Model Refinement**: Employing ensemble methods such as random forests or boosting to improve decision tree performance.
2. **Advanced Clustering Techniques**: Exploring density-based clustering algorithms (e.g., DBSCAN) to handle complex, non-linear cluster shapes.
3. **Cross-Dataset Validation**: Testing the models on diverse datasets to evaluate their generalizability across populations and settings.
4. **Integration with Clinical Tools**: Developing software solutions that integrate these models into diagnostic workflows, with real-time predictions and visual explanations for clinicians.
5. **Ethical and Fair AI**: Ensuring models are unbiased and generalizable, particularly in healthcare, where disparities can have life-or-death implications.

7. References

1. **World Health Organization (WHO)**
   World Health Organization. (2021). *Breast cancer*. Retrieved from https://www.who.int/news-room/fact-sheets/detail/breast-cancer
2. **Breast Cancer Wisconsin Dataset**
   Wolberg, William, Mangasarian, Olvi, Street, Nick, and Street, W. (1995). *Breast Cancer Wisconsin (Diagnostic)*. UCI Machine Learning Repository. Retrieved from https://doi.org/10.24432/C5DW2B
3. **Machine Learning in Breast Cancer Diagnostics**
   Delen, D., Walker, G., & Kadam, A. (2005). *Predicting breast cancer survivability: A comparison of three data mining methods*. Artificial Intelligence in Medicine, 34(2), 113-127.
4. **K-Nearest Neighbors in Breast Cancer Detection**
   Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques*. Elsevier.
5. **Clinical Applications of Machine Learning**
   Esteva, A., Robicquet, A., Ramsundar, B., et al. (2019). *A guide to deep learning in healthcare*. Nature Medicine, 25(1), 24–29.
6. **Importance of Data Preprocessing in ML**
   Kotsiantis, S. B., Kanellopoulos, D., & Pintelas, P. (2006). *Data preprocessing for supervised learning*. International Journal of Computer Science, 1(2), 111-117.
7. **Decision Tree Analysis in Healthcare**
   Quinlan, J. R. (1986). *Induction of decision trees*. Machine Learning, 1(1), 81-106.
8. **Logistic Regression in Medical Research**
   Hosmer, D. W., & Lemeshow, S. (2000). *Applied Logistic Regression*. Wiley.