CMPE479 – Final Project
Project Report

Breast Cancer Detection Using Machine Learning

Submitted by:

Yousef Khalil (CME) 20201701081

Project Mentor:

Assistant Prof. Rahim Dehkharghani

Faculty of Engineering and Natural Sciences

Kadir Has University

Fall 2023

TABLE OF CONTENTS

1 ABSTRACT

Breast cancer ranks as one of the most prevalent cancers worldwide and poses significant challenges in diagnosis and treatment. This project harnesses the power of data mining techniques to analyze a comprehensive breast cancer dataset, aiming to enhance diagnostic accuracy. Utilizing a dataset comprising various clinical measurements from 570 patients, the study applies normalization, visualization, and multiple data mining algorithms, including K-Means, Agglomerative Clustering, and Logistic Regression. The research focuses on identifying distinct patterns that differentiate between benign and malignant tumors and explores the relationships among various medical measurements. By comparing the effectiveness of different algorithms, this study sheds light on the potential of machine learning in medical diagnostics. The findings offer valuable insights for healthcare professionals and contribute to the ongoing research in oncological data analysis. This report presents the methodology, analysis, and conclusions drawn from the data, highlighting the project's significance in advancing breast cancer research and its implications for future diagnostic innovations.

2 INTRODUCTION

Breast cancer, one of the most common cancers affecting women worldwide, poses significant health challenges and necessitates early and accurate detection for effective treatment. The advent of data mining in healthcare has opened new avenues for enhancing diagnostic accuracy and personalized treatment strategies. This project focuses on a comprehensive analysis of a breast cancer dataset, utilizing advanced data mining techniques to extract valuable insights from clinical data. The dataset used in this study comprises various medical measurements associated with breast cancer tumors, it includes data on 570 patients, encompassing both benign and malignant cases. This dataset provides a foundation for applying various data mining methods to understand the patterns and characteristics of breast cancer.

2.1 Objective & Purpose

The primary objective of this breast cancer data mining project is to effectively distinguish between benign and malignant tumors using a dataset of 570 patients, enriched with a variety of medical measurements. This analysis aims not only to identify key indicative features and patterns for cancer diagnoses but also to explore the relationships among these measurements to gain a deeper understanding of breast cancer characteristics. The purpose of this attempt is to contribute significantly to breast cancer research, offering insights that could enhance diagnostic accuracy and inform treatment strategies. The project is expected to yield predictive models for breast cancer, shed light on critical factors influencing tumor malignancy, and provide a comparative analysis of various data mining techniques. Ultimately, this study aims to benefit healthcare professionals, researchers, and patients by advancing the knowledge of

breast cancer, potentially guiding targeted treatment approaches, and highlighting the pivotal role of machine learning and data mining in revolutionizing healthcare diagnostics, particularly in oncology.

3 METHODOLOGY

This section outlines the systematic approach adopted in this study to analyze the breast cancer dataset. The methodology encompasses several key stages: data collection, preprocessing, visualization, application of data mining techniques, and model evaluation.

3.1 Data Collection

The dataset, obtained from an online public repository, includes comprehensive medical records of 570 patients diagnosed with breast cancer. It features a range of clinical and pathological attributes pertinent to breast cancer diagnosis, ensuring a robust foundation for the analysis.

3.2 Data Preprocessing

Data preprocessing is a critical step in ensuring the quality and effectiveness of the data mining process. This stage involved:

- Removing Noise: Irrelevant or redundant data points were identified and removed to enhance the dataset's clarity and focus.

- Handling Missing Values: Missing data within the dataset were addressed either through imputation or exclusion, depending on the extent and nature of the missing information.
- Normalizing Numerical Attributes: Given the varying scales of the clinical measurements, normalization techniques were applied to bring all numerical attributes to a common scale, facilitating more accurate analysis and comparisons.

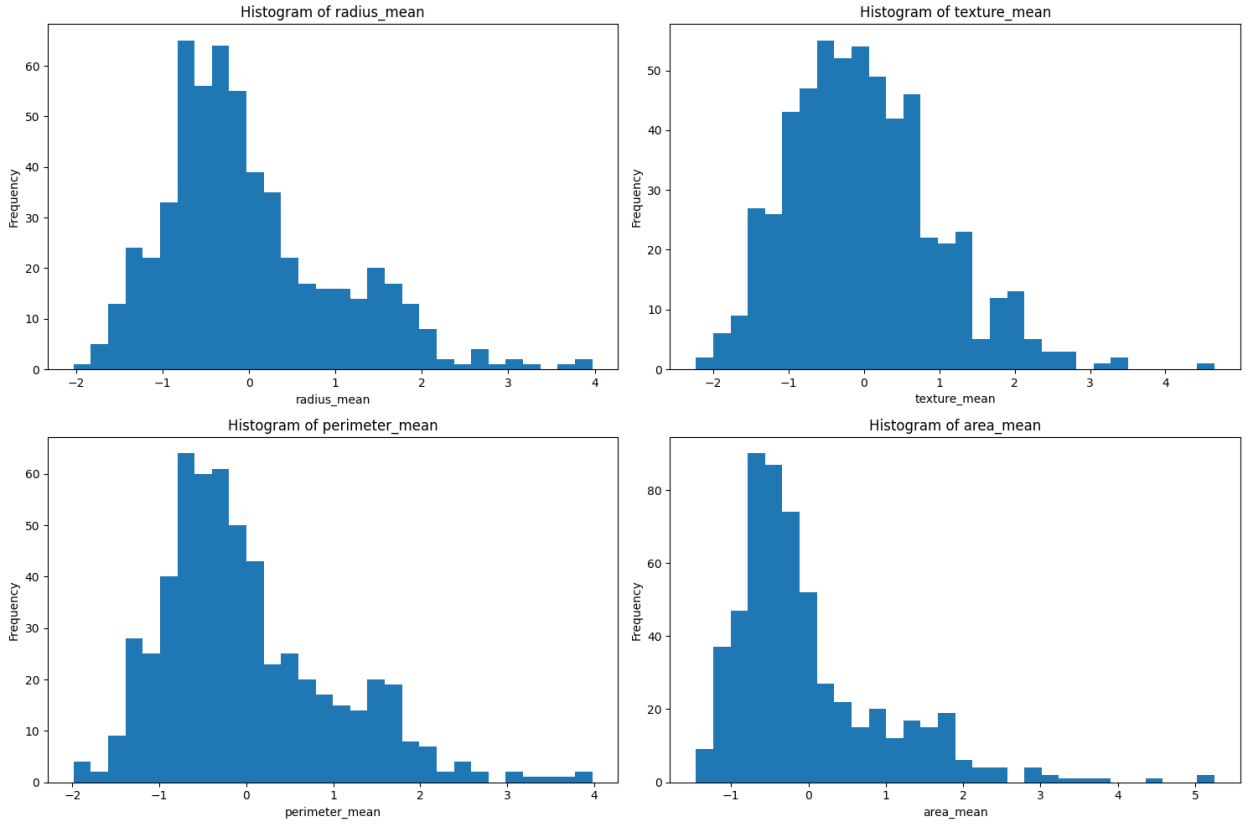
```python
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans

data  = pd.read_csv('C:\\Users\\youse\\Downloads\\cancer dataset.csv')

#data = data.drop('id',axis=1)

data['diagnosis'] = data['diagnosis'].map({'M': 1, 'B': 0})

# Check for missing values
missing_values = data.isnull().sum()
print("Missing values in each column:\n", missing_values)

# Check for duplicate rows
duplicate_rows = data.duplicated().sum()
print("Number of duplicate rows: ", duplicate_rows)

# Check for outliers in all columns
for column in data.columns:
    Q1 = data[column].quantile(0.25)
    Q3 = data[column].quantile(0.75)
    IQR = Q3 - Q1
    outliers = data[(data[column] < Q1 - 1.5 * IQR) | (data[column] > Q3 + 1.5 * IQR)]
    print(f"Number of outliers in {column}: ", len(outliers))


# Normalizing the data
scaler = StandardScaler()
# fit and transform the data
data_normalized = pd.DataFrame(scaler.fit_transform(data.drop('id', axis=1)), columns=data.columns.drop('id'))

# add the diagnosis column back to the normalized data
data_normalized['diagnosis'] = data['diagnosis']

# Find the value count of 1 (Malignant) and 0 (Benign) in the 'diagnosis' column
diagnosis_counts = data_normalized['diagnosis'].value_counts()

# Print the counts
print(diagnosis_counts)

data_normalized.head()
```

✓ 1.0s

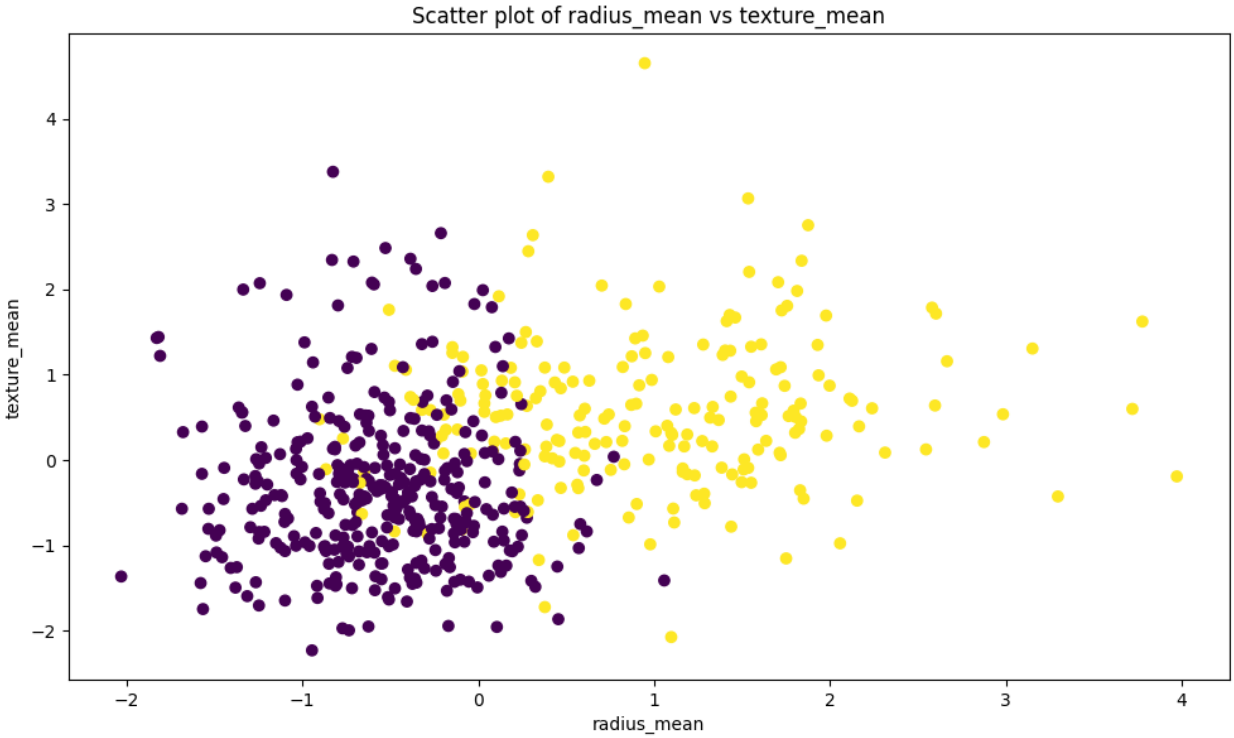
3.3 Data Visualization

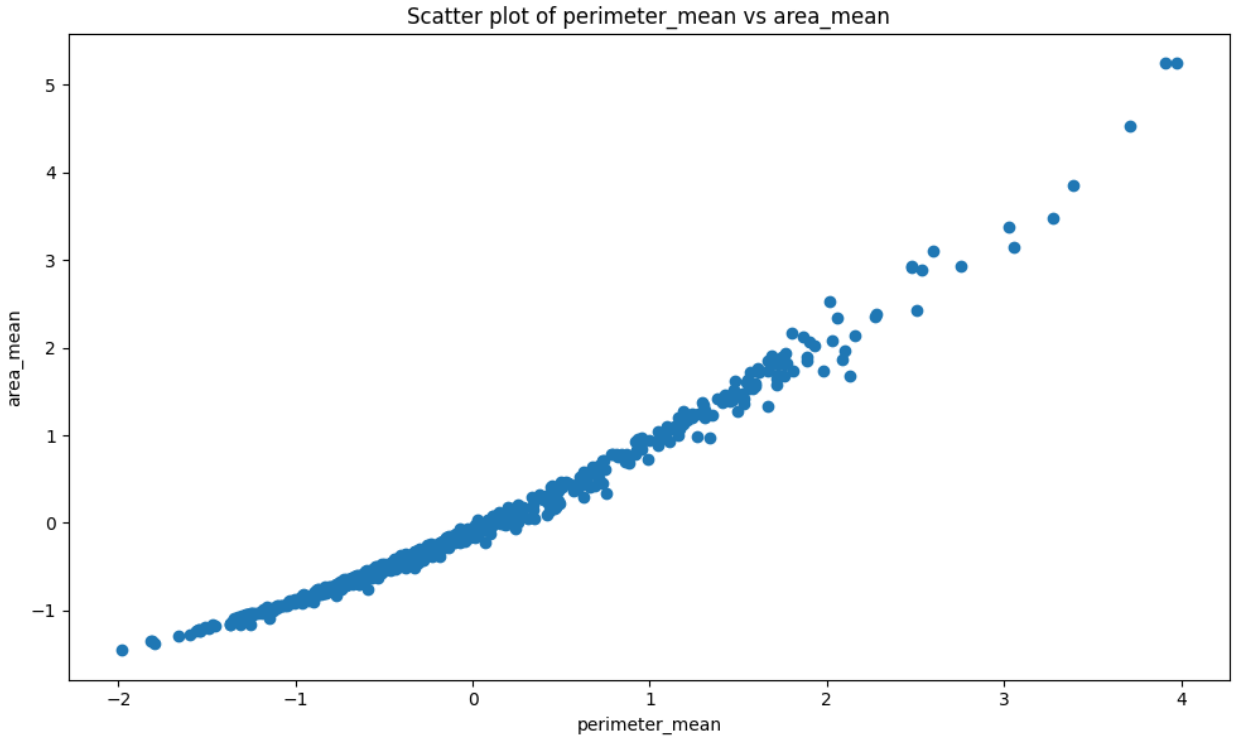
To understand the underlying patterns and relationships within the dataset, various visualization techniques were employed:

- Histograms: These were used to observe the distribution of different features within the dataset.

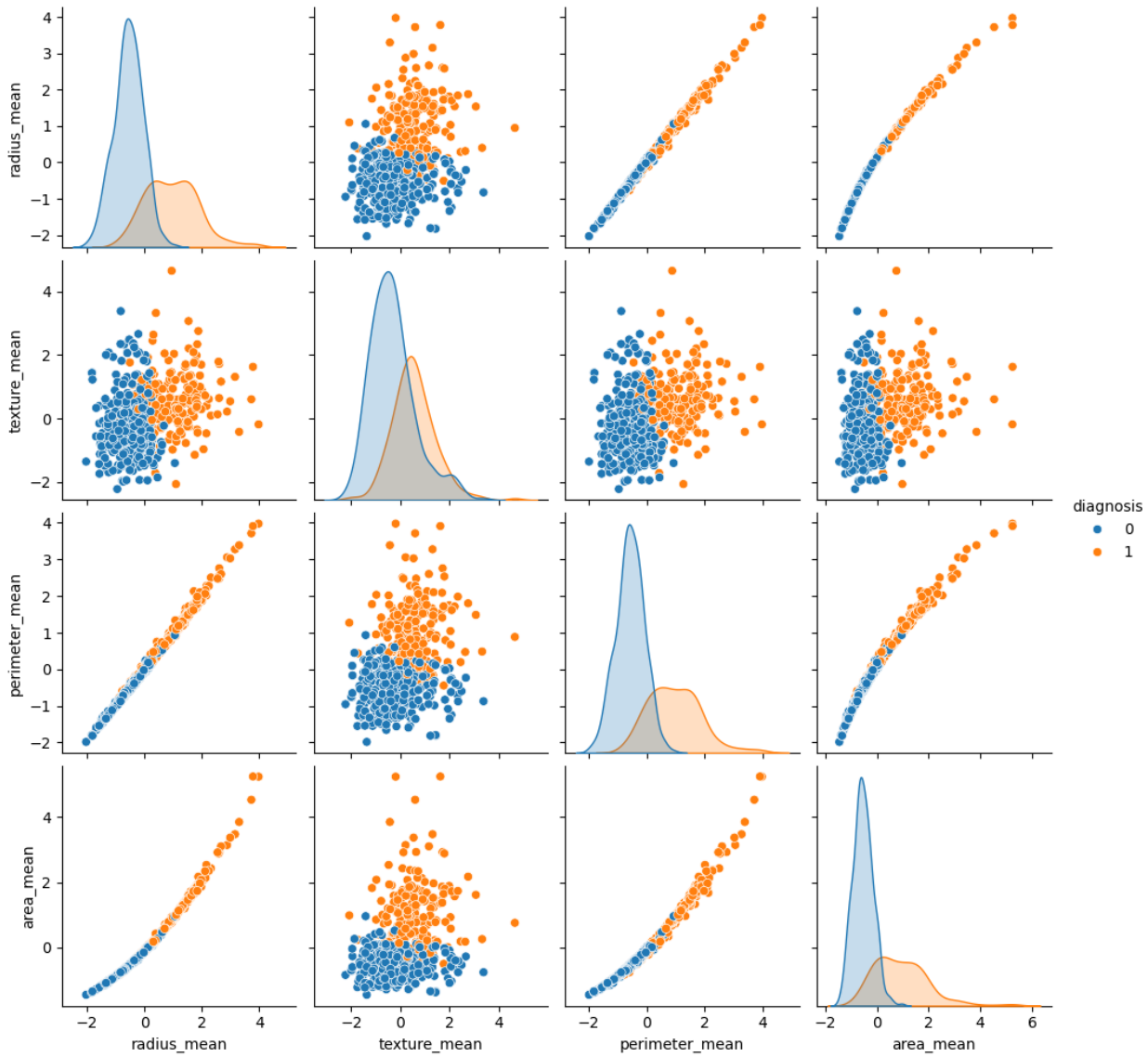


- Scatter Plots: Scatter plots provided insights into the relationships between pairs of features.

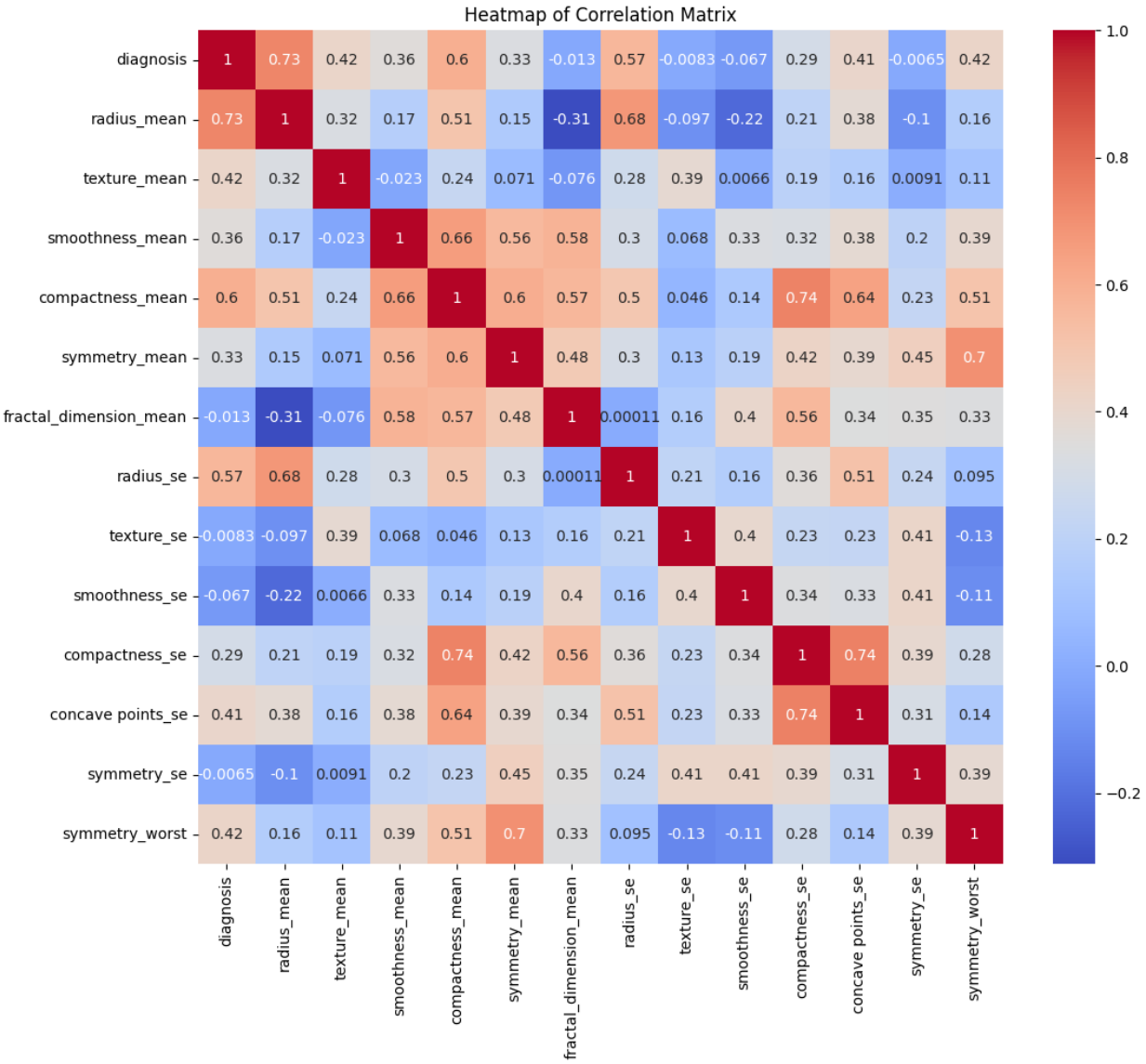

Scatter plot of radius_mean vs texture_mean
texture_mean
radius_mean
4
3
2
1
0
−1
−2
−2
−1
0
1
2
3
4


Scatter plot of perimeter_mean vs area_mean
area_mean
perimeter_mean
5
4
3
2
1
0
−1
−2
−1
0
1
2
3
4

- Pairplot: A pairplot was generated to visualize pairwise relationships across multiple dimensions within the dataset. The pairplot suggests that size-related features have a strong positive correlation with each other and can be indicative of the diagnosis. Features related to the size of the tumor (like `radius_mean`, `perimeter_mean`, and `area_mean`) are particularly distinct between benign and malignant tumors, which could be useful for classification purposes.



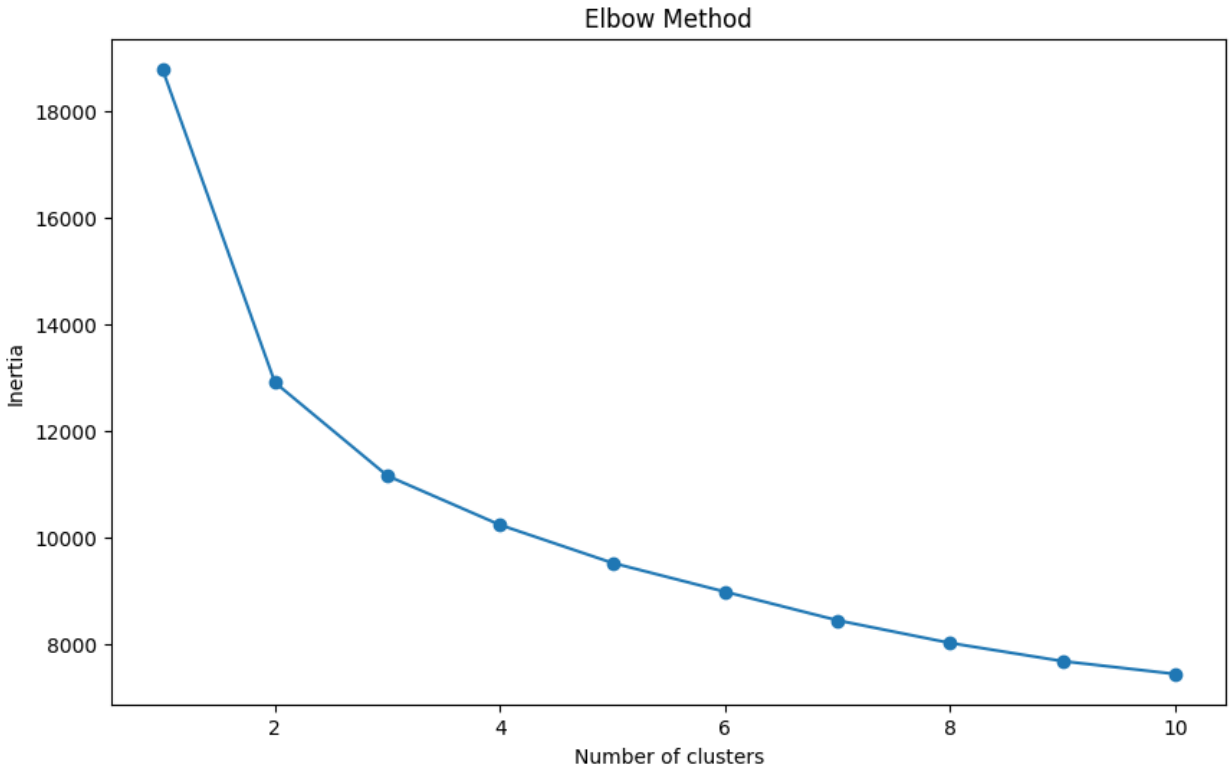
- Heatmap of Correlation Matrix: To understand the interdependencies between the variables, a heatmap was created using the correlation matrix. This visualization highlighted the strength and direction of the relationships between different features, identifying strongly, weakly, and negatively correlated quantities.



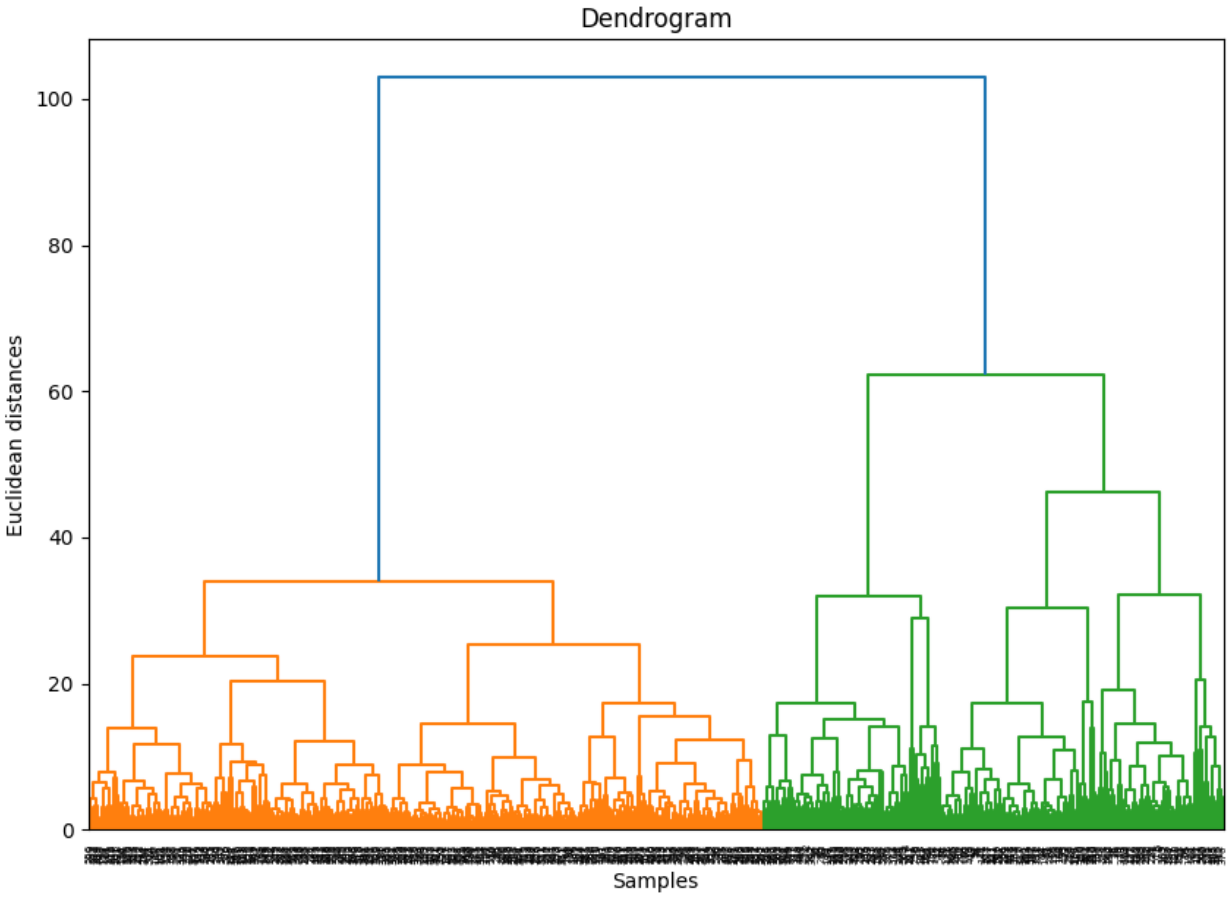
3.4 Data Mining Techniques

Multiple data mining algorithms were applied to analyze the dataset:

- K-Means Clustering: This unsupervised learning technique was used to group the data into clusters based on feature similarities.



- Agglomerative Clustering: Another clustering technique, based on hierarchical clustering principles, was used to group data points and visualize them in a dendrogram.



- Logistic Regression: As a supervised learning approach, logistic regression was employed to model the probability of binary outcomes (benign or malignant tumors).

3.5 Model Evaluation

The effectiveness of the applied models was evaluated through several metrics:

- Accuracy and Confusion Matrices: These metrics were crucial in assessing the performance of the Logistic Regression model, especially in terms of its ability to correctly classify the tumor types.
- Train-Test Split: The dataset was split into training and testing sets (70/30) to validate the model's performance on unseen data.

4 RESULTS

This section discusses the results obtained from the application of different data mining techniques on the breast cancer dataset. The study focused on comparing the effectiveness of K-Means clustering, Agglomerative Clustering, and Logistic Regression. Key metrics were used to assess the performance of these models.

4.1 K-Means Clustering

- Silhouette Score: The K-Means algorithm achieved a silhouette score of 0.3274, indicating a moderate separation between clusters. This score suggests that the clusters are not highly dense and well-separated but still offer a reasonable structure within the data.

- Davies-Bouldin Score: The Davies-Bouldin score of 1.4752 for K-Means suggests that the clusters are somewhat average in terms of similarity and separation. A lower score is typically desired, indicating better clustering.

4.2 Agglomerative Clustering

- Silhouette Score: Agglomerative Clustering resulted in a silhouette score of 0.2788, which is slightly lower than that of K-Means. This score reflects less distinction between clusters compared to K-Means.
- Davies-Bouldin Score: With a Davies-Bouldin score of 1.6193, Agglomerative Clustering indicates a lesser degree of separation between clusters compared to K-Means.

4.3 Logistic Regression

- Training Accuracy: The Logistic Regression model showed a high training accuracy of 98.74%, which demonstrates its effectiveness in classifying the data correctly.
- Test Accuracy and Confusion Matrix: The model achieved a perfect accuracy of 100% on the test set. The confusion matrix showed that all 108 benign (negative) and 63 malignant (positive) cases were correctly classified with no false positives or negatives.

4.4 Comparative Analysis

- Performance Comparison: Among the clustering techniques, K-Means exhibited relatively better performance in terms of both silhouette and Davies-Bouldin scores, suggesting more distinct and well-separated clustering than Agglomerative Clustering.
- Model Suitability: Logistic Regression outperformed the clustering techniques in terms of predictive accuracy. The perfect accuracy on the test set, while impressive, should be cautiously interpreted, as it might also indicate overfitting. However, the high training accuracy supports the model's robustness.
- Implications: The differences in the performance of K-Means and Agglomerative Clustering highlight the varying suitability of clustering algorithms based on the dataset's characteristics. Logistic Regression’s high accuracy underscores its potential for predictive tasks in binary classification problems like cancer diagnosis.

4.5 Discussion

- The results demonstrate the efficacy of Logistic Regression in classification tasks, particularly in medical datasets where accurate diagnosis is crucial. The clustering techniques, while not as precise as Logistic Regression for direct classification, provided valuable insights into the dataset's structure. K-Means, with its moderate silhouette score, proved useful for initial exploratory analysis. However, the perfect accuracy in Logistic Regression’s test set results warrants further investigation to rule out any overfitting and confirm the model’s generalizability.

- These findings contribute to the broader understanding of applying data mining techniques in medical diagnostics and emphasize the importance of choosing appropriate algorithms based on the specific nature and requirements of the dataset.

5 CONCLUSION

This report presented a comprehensive data mining analysis of a breast cancer dataset, utilizing various statistical and machine learning techniques to draw meaningful insights into the diagnosis and characteristics of breast cancer. Through the methodologies employed, including data normalization, visualization, clustering, classification, and model evaluation, the study aimed to enhance the understanding of breast cancer's diagnostic patterns.

The visualizations, including histograms, scatter plots, pairplots, and a heatmap, revealed critical insights into the dataset. Size-related features like radius_mean, perimeter_mean, and area_mean demonstrated strong positive correlations and distinct distributions between benign and malignant tumors, indicating their potential as significant predictors in classification models.

In applying data mining techniques, K-Means and Agglomerative Clustering provided valuable exploratory analysis, with K-Means showing relatively better performance in cluster separation as indicated by its silhouette and Davies-Bouldin scores. However, the Logistic Regression model stood out in its predictive ability, achieving high accuracy in classifying breast cancer tumors. The perfect test set accuracy, while impressive, suggests the need for further validation to ensure the model's generalizability.

The results underscore the importance of selecting appropriate data mining techniques based on the dataset characteristics. Logistic Regression's success in this study highlights its suitability for binary classification problems in medical diagnostics. However, the clustering methods' insights into the data structure demonstrate the value of exploratory analysis in understanding complex medical datasets.

In conclusion, this project not only demonstrated the effectiveness of various data mining techniques in analyzing medical data but also emphasized the potential of machine learning in enhancing diagnostic accuracy in healthcare.