BREAST CANCER CLUSTERING PROJECT REPORT

STUDENT NAME: RITIKA SEN
COURSE: COMPUTER SCIENCE ENGINEERING (DATA SCIENCE)

INSTITUTE: HERITAGE INSTITUTE OF TECHNOLOGY

Period of Internship: 25th August 2025 - 19th September 2025

Report submitted to: IDEAS – Institute of Data Engineering, Analytics and Science Foundation, ISI

Kolkata

Project Report

1. Abstract

This project explores unsupervised learning on the Breast Cancer dataset. The dataset contains 30 medical features related to cell nuclei from breast cancer biopsies. The main goal was to analyze high-dimensional data, reduce its dimensionality using Principal Component Analysis (PCA), and then apply KMeans clustering to identify natural groupings in the data. Exploratory Data Analysis (EDA) was performed to understand feature relationships through correlation heatmaps and visualizations. PCA reduced the dataset to two dimensions, enabling easy visualization of clusters. KMeans clustering was applied to group patients, and silhouette scores were used to evaluate performance. The project demonstrates the usefulness of dimensionality reduction and clustering in medical data analysis.

2. Introduction

Breast cancer is one of the most common types of cancer in the world, and early detection can improve survival rates significantly. This project leverages machine learning techniques to analyze the Breast Cancer dataset, a high-dimensional dataset with 30 medical attributes.

Relevance: Applying unsupervised learning can uncover patterns in medical data without prior labels, supporting exploratory insights.

Technology Used: Python, Scikit-learn, Matplotlib, Seaborn, PCA, KMeans.

Procedure:
1. Load and explore dataset features.
2. Perform data preprocessing and scaling.
3. Use PCA to reduce feature dimensions to 2.
4. Apply KMeans clustering on reduced data.
5. Visualize and evaluate results.

Purpose: To demonstrate how high-dimensional data can be simplified for visualization and clustering using PCA and KMeans.

3. Project Objective

- To explore and understand high-dimensional medical data through EDA.
- To apply PCA for dimensionality reduction of the dataset from 30 to 2 dimensions.
- To perform KMeans clustering to identify natural groupings in the data.
- To evaluate clustering performance using silhouette score.
- To visualize clusters using Matplotlib and OpenCV.

4. Methodology

1. Data Loading: The Breast Cancer dataset was loaded from sklearn.datasets. It contains 569 samples with 30 features.
2. Exploratory Data Analysis (EDA): Correlation heatmap plotted to see feature relationships, scatter plots for mean radius, mean texture, mean perimeter, etc.
3. Preprocessing: Data was scaled using StandardScaler to normalize features.
4. Dimensionality Reduction: PCA was applied to reduce dimensions from 30 to 2.
5. Clustering: KMeans clustering was performed with 2–3 clusters, and cluster labels were assigned.
6. Evaluation: Silhouette score was calculated to evaluate cluster quality.
7. Visualization: PCA scatter plot with KMeans centroids using Matplotlib and OpenCV visualization.

5. Data Analysis and Results

- Descriptive Analysis:
  The dataset had 30 numerical features, all continuous. Correlation heatmap revealed strong correlation between mean radius, mean perimeter, and mean area.
- Inferential Analysis:
  PCA reduced the data to 2D while preserving ~95% variance. Clusters were formed around benign and malignant samples.
- Results:
  KMeans clustering separated the dataset into clusters with reasonable overlap to true labels. Silhouette score indicated moderate cluster quality.
- Visualizations:
  Correlation heatmap, PCA scatter plots with cluster centers, OpenCV cluster visualization.

6. Conclusion

The project successfully demonstrated how PCA and KMeans can be combined to handle high-dimensional medical data. PCA allowed easier visualization and reduced computational complexity, while KMeans provided an unsupervised way to group patients into clusters. Though clustering is not perfect, the results aligned reasonably well with the malignant vs benign separation.

Recommendations for future work:
- Apply supervised models (like Logistic Regression, Random Forest, SVM) to achieve higher prediction accuracy.
- Experiment with other clustering methods like DBSCAN or hierarchical clustering.
- Use larger real-world medical datasets for better generalization.

7. Appendices

1. References:
   - Scikit-learn Documentation
   - Breast Cancer Dataset (UCI ML Repository)
2. Github Link: https://github.com/RitikaSen27/IDEAS-TIH-INTERNSHIP.git
3. Visualizations: Screenshots of correlation heatmaps, PCA scatter plots, and OpenCV visualizations.