A COMPARATIVE ANALYSIS OF VARIOUS CLUSTERING TECHNIQUES ON DIFFERENT DATASETS WITH PRIVACY PRESERVATION

19CSPN6601- INNOVATIVE AND CREATIVE PROJECT

Submitted by

KRISHNA PRASATH R	**727621BCS034**
HARINI P	**727621BCS046**
SANOFER NISWAN S	**727621BCS064**

in partial fulfillment for the award of the degree
of

Bachelor of Engineering

in

Computer Science and Engineering

Dr. Mahalingam College of Engineering and Technology

Pollachi - 642003

(An Autonomous Institution Affiliated to Anna University, Chennai)

MAY 2024

Dr. MAHALINGAM COLLEGE OF ENGINEERING AND TECHNOLOGY, POLLACHI -642003

(An Autonomous Institution Affiliated to Anna University, Chennai)

BONAFIDE CERTIFICATE

Certified that this Project Report, "A COMPARATIVE ANALYSIS OF VARIOUS CLUSTERING TECHNIQUES ON DIFFERENT DATASETS WITH PRIVACY PRESERVATION"

is the bonafide work of

KRISHNA PRASATH R	727621BCS034
HARINI P	727621BCS046
SANOFER NISWAN S	727621BCS064

who carried out the Innovative and Creative project work under my supervision.

Ms. C. Devipriya
SUPERVISOR
ASSISTANT PROFESSOR (SS)
Computer Science and Engineering
Dr.Mahalingam College of Engineering
Technology,Pollachi- 642003

Dr. G. Anupriya
HEAD OF THE DEPARTMENT
PROFESSOR
Computer Science and Engineering
Dr. Mahalingam College of Engineering and
Technology, Pollachi – 642003

Submitted for the Autonomous End Semester Examination Innovative and Creative Project Viva-voce held on ______________________

INTERNAL EXAMINER **EXTERNAL EXAMINER**

Dr. Mahalingam College of Engineering and Technology
Pollachi -642003

Technology Readiness Level (TRL) Certificate

Project Title: A COMPARATIVE ANALYSIS OF VARIOUS CLUSTERING TECHNIQUES ON DIFFERENT DATASETS WITH PRIVACY PRESERVATION

Course Code: 19CSPN6601

Students Names and Roll Numbers:

KRISHNA PRASATH R	727621BCS034
HARINI P	727621BCS046
SANOFER NISWAN S	727621BCS064

Guide Name: Ms.C.Devipriya AP(SS)/CSE

Technology Readiness Level (TRL) of this Project: ___________________

Signature of the Guide HoD

Internal Examiner External Examiner

A COMPARATIVE ANALYSIS OF VARIOUS CLUSTERING TECHNIQUES ON DIFFERENT DATASETS WITH PRIVACY PRESERVATION

ABSTRACT

To enhance dataset privacy, our project integrates an investigation into how privacy preservation techniques impact clustering algorithm performance. By assessing how various methods affect clustering outcomes, we aim to elucidate the trade-offs between privacy and clustering quality. Differential privacy, among other techniques, is applied to datasets prior to clustering analysis. Initially, common algorithms like k-means, hierarchical, and spectral clustering are employed without privacy measures. Subsequently, differential privacy is introduced, and the same algorithms are applied to modified datasets. Performance metrics, including clustering quality and scalability, are assessed before and after privacy application. This analysis serves a critical purpose, ensuring that while protecting sensitive data, meaningful patterns can still be extracted. Understanding these trade-offs is paramount for adhering to privacy regulations and ethical considerations. By investigating how privacy preservation affects clustering outcomes, we contribute to a deeper understanding of privacy's impact on analysis. This knowledge empowers practitioners to make informed decisions about balancing privacy concerns with analytical goals, advancing both data privacy and analytical accuracy.

ACKNOWLEDGEMENT

First and foremost, we wish to express our deep unfathomable feeling, gratitude to our institution and our department for providing us a chance to fulfill our long cherished dreams of becoming Computer Science Engineers.

We express our sincere thanks to our honorable Secretary **Dr.C.Ramaswamy** for providing us with required amenities.

We wish to express our hearty thanks to **Dr.P.Govindasamy,** Principal of our college, for his constant motivation and continual encouragement regarding our innovative and creative project work.

We are grateful to **Dr.G.Anupriya,** Head of the Department, Computer Science and Engineering, for her direction delivered at all times required. We also thank her for her tireless and meticulous efforts in bringing out this innovative and creative project to its logical conclusion.

Our hearty thanks to our guide **Ms.C.Devipriya,** Assistant Professor (SS),for her constant support and guidance offered to us during the course of our project by being one among us and all the noble hearts that gave us immense encouragement towards the completion of our innovative and creative project.

We also thank our review panel members for their continuous support and guidance.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CHS	Calinski-Harabasz Score
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DB Score	Davies-Bouldin Score
EPS	Epsilon (Parameter used in DIfferential Privacy)
HC	Hierarchical Clustering
MAE	Mean Absolute Error
MSE	Mean Squared Error
n_clusters	Number of Clusters
PCA	Principal Component Analysis
PPDM	Privacy Preserving Data Mining
SC	Spectral Clustering
WCSS	Within-Cluster Sum of Squares

CHAPTER 1

INTRODUCTION

CHAPTER 1

INTRODUCTION

In an era of increasing data digitization and utilization, ensuring the privacy of sensitive information has become paramount. With the proliferation of data-driven applications, including healthcare, finance, and e-commerce, the need to safeguard personal and confidential data has never been more critical. Clustering, a fundamental technique in data analysis, plays a pivotal role in identifying meaningful patterns and structures within datasets. However, the application of clustering algorithms to sensitive data poses inherent privacy risks, as it may inadvertently reveal personal information about individuals. To address this challenge, privacy-preserving techniques, such as differential privacy, have emerged as promising solutions to protect the confidentiality of data while still enabling meaningful analysis.

This project aims to evaluate the effectiveness of privacy-preserving techniques, particularly differential privacy, in safeguarding sensitive data during the clustering process. By applying differential privacy mechanisms, such as noise addition and data distortion, we assess the extent to which privacy measures can mitigate the risk of unauthorized disclosure of personal information. Furthermore, we investigate the impact of these privacy measures on the accuracy of clustering algorithms, considering factors such as clustering performance and data utility.

Central to our analysis is the exploration of the trade-offs between privacy preservation and clustering accuracy. We seek to understand the delicate balance between ensuring data privacy and maintaining the quality and reliability of clustering results. By examining the interplay between privacy measures, clustering algorithms, and data utility, we aim to provide insights that inform decision-making in real-world scenarios. Ultimately, this project contributes to the broader discourse on privacy-preserving data analysis, offering valuable implications for industries and organizations grappling with the dual imperatives of data privacy and analytical utility.

1.1 DOMAIN - MACHINE LEARNING AND DATASCIENCE

In the field of data science and Machine Learning, this project explores the intersection of clustering techniques and privacy algorithms. With a focus on real-world datasets, the study aims to assess the efficacy of various clustering methods in maintaining data privacy while retaining analytical accuracy.

1.2 OBJECTIVE OF THE PROJECT

Evaluate Privacy Measures: Assess the effectiveness of privacy-preserving techniques, like differential privacy, in safeguarding sensitive data during clustering. Analyze Clustering Accuracy: Examine how different clustering algorithms perform after applying privacy measures, considering factors like noise addition and data distortion. Assess Data Utility: Evaluate the utility of data after dimensionality reduction using PCA and its impact on clustering performance and privacy preservation.

1.3 PROBLEM STATEMENT

In today's data-driven landscape, organizations face the challenge of reconciling the demand for data-driven insights with the imperative to protect individuals' privacy. Clustering, a fundamental technique in data analysis, poses particular risks to privacy when applied to sensitive datasets. Traditional clustering algorithms may inadvertently disclose personal information, raising concerns about privacy violations and regulatory compliance. To address this issue, there is a growing need for effective privacy-preserving techniques that can safeguard sensitive data while enabling meaningful analysis. However, the effectiveness of such techniques in balancing privacy preservation and clustering accuracy remains an open question. This project seeks to investigate the efficacy of privacy-preserving methods, specifically differential privacy, in mitigating privacy risks during clustering. By evaluating the impact of differential privacy mechanisms on clustering accuracy and data utility, this research aims to provide insights into the trade-offs between privacy protection and analytical utility. Ultimately, the findings will inform decision-making in industries and organizations striving to navigate the complex landscape of data privacy and analysis.

CHAPTER 2

LITERATURE SURVEY

CHAPTER 2

LITERATURE SURVEY

2.1 A Comparative Study of Clustering Algorithms

Author: Manoj Kr Gupta and Pravin Chandra

Manoj et al. offers a comparative analysis of clustering algorithms like K-means, DBSCAN, hierarchical clustering, and spectral clustering. It examines their performance across diverse datasets, focusing on clustering quality, scalability, and robustness. The authors' insights aid practitioners in selecting suitable techniques based on dataset characteristics. Leveraging these insights, our project aims to understand how privacy preservation techniques affect clustering algorithms across datasets, refining our approach to balancing privacy and accuracy. These findings will guide our selection of clustering methods for a comprehensive evaluation under privacy-preserving conditions.

2.2 Hierarchical Clustering Algorithms: An Overview

Author: Murtagh, Fionn, and Pedro Contreras

This study provides an insightful overview of hierarchical clustering algorithms, elucidating the principles and methodologies behind this clustering approach. It explores various techniques employed in hierarchical clustering, including agglomerative and divisive methods, and discusses their applications in data mining and knowledge discovery. By examining the advantages and limitations of hierarchical clustering, the authors offer valuable guidance for researchers and practitioners seeking to leverage this method for data analysis tasks. Leveraging these insights, our project aims to incorporate hierarchical clustering as a fundamental technique in our comparative analysis, evaluating its performance under privacy-preserving conditions across diverse datasets. The nuanced understanding provided by Murtagh and Contreras will inform our assessment of hierarchical clustering's suitability for different types of data and its implications for privacy preservation in clustering applications analysis.

2.3 Big Data Privacy Preservation Using Principal Component Analysis and Random Projection in Healthcare

Author: Ratra, Ritu

Ratra and colleagues delve into the realm of privacy preservation in healthcare big data through innovative techniques like principal component analysis (PCA) and random projection. They propose a framework aimed at safeguarding sensitive healthcare information while ensuring data utility for analysis. By employing PCA and random projection, the authors offer a pragmatic solution to the challenges of privacy preservation in the context of healthcare data analytics. Leveraging these techniques, our project aims to explore effective privacy-preserving strategies in healthcare data clustering. The insights from Ratra et al.'s work will guide our evaluation of PCA and random projection methods in maintaining data privacy while retaining analytical efficacy, contributing to the advancement of privacy-preserving techniques in healthcare analytics.

2.4 A Comprehensive Review on Privacy Preserving Data Mining

Authors: Aldeen, Yousra Abdul Alsahib S., Mazleena Salleh, and Mohammad Abdur Razzaque

Aldeen et al. provide an in-depth examination of privacy-preserving data mining techniques. They analyze various approaches and methodologies aimed at preserving the privacy of sensitive data while allowing for meaningful analysis. Through their critical review, the authors offer insights into the challenges and advancements in the field of privacy-preserving data mining. Leveraging the findings from Aldeen et al.'s work, our project aims to gain a deeper understanding of privacy-preserving techniques and their implications for data clustering. Their review will inform our selection and evaluation of privacy-preserving methods in our comparative analysis, contributing to the enhancement of privacy in data mining applications.

2.5 Non-linear Dimensionality Reduction for Privacy-Preserving Data Classification

Authors: Alotaibi, Khaled

Alotaibi et al. explore non-linear dimensionality reduction techniques for privacy-preserving data classification. They investigate methods that can effectively reduce the dimensionality of data while preserving privacy. Leveraging the insights from Alotaibi et al.'s work, our project seeks to evaluate the effectiveness of non-linear dimensionality reduction techniques in preserving privacy during data clustering tasks. Their findings will guide our assessment of privacy-preserving methods in clustering analysis, facilitating the development of more secure and accurate data mining approaches.

2.6 Use of Principal Component Analysis (PCA) and Hierarchical Cluster Analysis (HCA) for Multivariate Association between Bioactive Compounds and Functional Properties in Foods: A Critical Perspective

Authors: Granato, Daniel

Granato et al. offer a critical perspective on the utilization of principal component analysis (PCA) and hierarchical cluster analysis (HCA) in assessing the multivariate association between bioactive compounds and functional properties in foods. Through their analysis, the authors provide insights into the strengths and limitations of PCA and HCA in elucidating complex relationships in food composition data. Leveraging the insights from Granato et al.'s work, our project aims to explore the application of PCA and HCA in clustering analysis, particularly in the context of privacy preservation. Their critical perspective will inform our evaluation of these techniques' efficacy in maintaining data privacy while extracting meaningful patterns from multivariate datasets.

2.7 Privacy Preserving Clustering

Authors: Jha, Somesh, Luis Kruger, and Patrick McDaniel

Jha, Kruger, and McDaniel delve into the realm of privacy-preserving clustering, offering insights into techniques aimed at maintaining data privacy during clustering analysis. Their work, presented at the 10th European Symposium on Research in Computer Security, explores methodologies for ensuring confidentiality while extracting meaningful clusters from sensitive data. Leveraging the findings from Jha et al.'s research, our project aims to evaluate the effectiveness of privacy-preserving clustering methods in safeguarding sensitive information across various datasets. Their contributions will inform our exploration of privacy-preserving techniques and their implications for clustering accuracy and privacy preservation.

2.8 PCA-Based Feature Selection Scheme for Machine Defect Classification

Authors: Malhi, Arnaz, and Robert X. Gao

Malhi and Gao propose a feature selection scheme based on principal component analysis (PCA) for machine defect classification. Their work, published in the IEEE Transactions on Instrumentation and Measurement, focuses on enhancing the accuracy of defect classification by selecting the most relevant features through PCA. Leveraging their approach, our project aims to explore the application of PCA-based feature selection in clustering analysis, particularly in the context of privacy preservation. Their methodology will guide our evaluation of feature selection techniques for improving clustering accuracy while maintaining data privacy.

2.9 Research on Spectral Clustering Algorithms and Prospects

Authors: Ding, Shifei, Liwen Zhang, and Yu Zhang

Ding, Zhang, and Zhang present research on spectral clustering algorithms and their prospects at the 2010 2nd International Conference on Computer Engineering and Technology. Their work explores advancements in spectral clustering techniques, highlighting their potential for effectively partitioning data into meaningful clusters. Leveraging their insights, our project aims to evaluate the performance of spectral clustering algorithms in diverse clustering scenarios, including privacy-preserving clustering tasks. Their research will inform our selection and evaluation of spectral clustering methods, contributing to a comprehensive understanding of their capabilities and limitations in clustering analysis.

2.10 Deep Clustering: Advances, Challenges, and Future Directions

Authors: Ren, Y., Pu, J., Yang, Z., Xu, J., Li, G., Pu, X.,& He, L.

Ren et al. provide a comprehensive survey on deep clustering, exploring its advances, challenges, and future directions. Their work, presented as an arXiv preprint, delves into the intersection of deep learning and clustering analysis, highlighting the potential of deep clustering for discovering complex patterns in high-dimensional data. Leveraging their insights, our project aims to investigate the application of deep clustering techniques in privacy-preserving clustering tasks. Their survey will inform our exploration of deep clustering algorithms and their implications for maintaining data privacy while achieving clustering accuracy.

2.11 Clustering Techniques for Streaming Data: A Survey

Author: Toshniwal, Durga

Toshniwal et al. explores clustering techniques tailored for streaming data analysis. The study investigates methodologies for clustering data streams in real-time, addressing the challenges posed by continuous data arrival and evolving data distributions. Leveraging insights from Toshniwal's survey, our project aims to evaluate the effectiveness of streaming data clustering techniques in privacy-preserving scenarios. Their comprehensive overview will inform our selection and assessment of clustering algorithms suitable for dynamic data environments, contributing to the advancement of privacy-preserving techniques in streaming data analysis.

2.12 Dynamic Clustering

Author: Bouchachia, Abdelhamid

Bouchachia's work on dynamic clustering, published in Evolving Systems, delves into the intricacies of clustering algorithms designed to adapt to evolving data distributions. The study explores methodologies for dynamically partitioning data into clusters in response to changes in data characteristics over time. Leveraging insights from Bouchachia's research, our project aims to investigate dynamic clustering techniques' suitability for privacy-preserving clustering tasks. Their exploration of dynamic clustering methods will inform our evaluation of algorithms capable of maintaining data privacy while accommodating shifting data patterns, enhancing the applicability of clustering in dynamic environments.

2.13 Customer Segmentation in User Behavior Analysis: A Comparative Study of Clustering Algorithms

Author: Liu, Yingze

Liu et al. conducts a comparative study on clustering algorithms for customer segmentation in user behavior analysis, published in Highlights in Business, Economics and Management. The study evaluates the effectiveness of various clustering techniques in segmenting customers based on their behavioral patterns. Through rigorous analysis, Liu offers insights into the strengths and limitations of different clustering algorithms in identifying meaningful customer segments. Leveraging Liu's comparative study, our project aims to assess the applicability of clustering algorithms in privacy-preserving customer segmentation tasks. Their findings will guide our selection and evaluation of clustering methods for effectively segmenting customers while preserving their privacy, contributing to more targeted marketing strategies.

2.14 Comparative Analysis of the Applicability of Five Clustering Algorithms for Market Segmentation

Authors: Teslenko D, Sorokina A, Smelyakov K, et al.

Teslenko et al. present a comparative analysis of five clustering algorithms for market segmentation, presented at the 2023 IEEE Open Conference of Electrical, Electronic and Information Sciences. The study assesses the suitability of clustering algorithms for segmenting market data into meaningful groups. Through their analysis, the authors provide insights into the performance and applicability of various clustering techniques in market segmentation tasks. Leveraging Teslenko et al.'s comparative analysis, our project aims to evaluate clustering algorithms' effectiveness in privacy-preserving market segmentation. Their findings will inform our selection and assessment of clustering methods for segmenting markets while preserving data privacy, contributing to more accurate market analysis and targeted marketing strategies.

2.15 Two-Step Clustering for Data Reduction Combining DBSCAN and K-means Clustering

Authors: Kremers, B.J., Citrin, J., Ho, A., and van de Plassche, K.L.

Kremers et al. proposed a two-step clustering approach for data reduction that combines DBSCAN and k-means clustering techniques, published in Contributions to Plasma Physics. The study introduces a novel methodology for efficiently reducing the dimensionality of data while preserving its essential characteristics. Through their approach, the authors aim to improve clustering performance and scalability in high-dimensional datasets. Leveraging Kremers et al.'s methodology, our project aims to explore innovative techniques for privacy-preserving data clustering and reduction.

2.16 Summary

The literature survey provides a comprehensive overview of various clustering algorithms and techniques, focusing on their applications, advantages, and limitations, with a specific emphasis on privacy preservation. The studies discussed cover a wide range of clustering methodologies, including traditional approaches like K-means, hierarchical clustering, and spectral clustering, as well as more advanced techniques such as deep clustering and dynamic clustering. Additionally, several studies explore the integration of dimensionality reduction methods like principal component analysis (PCA) and feature selection schemes to enhance clustering performance while maintaining data privacy. Furthermore, the survey includes research on privacy-preserving clustering in specific domains such as healthcare data analytics and market segmentation.

The first set of studies offers a comparative analysis of clustering algorithms across diverse datasets, highlighting their performance in terms of clustering quality, scalability, and robustness. These insights aid in selecting suitable techniques based on dataset characteristics and inform the evaluation of clustering methods under privacy-preserving conditions.

Another group of studies focuses on specific clustering methodologies, such as hierarchical clustering and spectral clustering, elucidating their principles, methodologies, and applications. These insights contribute to a nuanced understanding of these techniques' capabilities and limitations in various clustering scenarios, including privacy-preserving clustering tasks.

Additionally, the survey includes research on innovative privacy-preserving techniques, such as PCA and random projection, aimed at safeguarding sensitive data while ensuring data utility for analysis. These techniques offer pragmatic solutions to the challenges of privacy preservation in the context of healthcare data analytics and other domains.

Furthermore, the survey explores the intersection of clustering with other domains, such as feature selection for machine defect classification and customer segmentation in user behavior analysis.Overall, the literature survey offers a comprehensive overview of clustering algorithms, techniques, and applications, with a specific focus on privacy preservation. The insights and methodologies discussed in these studies will inform the development of more secure and accurate privacy-preserving clustering approaches, contributing to the advancement of data mining and analytics in various domains.

CHAPTER 3

EXISTING SYSTEM

CHAPTER 3

EXISTING SYSTEM

3.1 OVERVIEW

The study emphasizes the significance of understanding customer behavior patterns in today's digital landscape for corporate success. Employing three distinct clustering algorithms - k-means, hierarchical clustering, and DBSCAN - the study aims to delve deeply into client segmentation. By meticulously analyzing factors such as age, yearly income, and consumption score, the study offers a comprehensive perspective on various consumer attributes using data from the mall consumer Segmentation Dataset. This insight serves as a valuable tool for adjusting marketing strategies, empowering stakeholders to make informed decisions and improve market performance. Additionally, the study maps the path to greater competitiveness and relevance in a developing market segment through the utilization of real-world data and powerful clustering techniques, demonstrating the efficacy of these algorithms in modern business environments.

3.2 BLOCK DIAGRAM

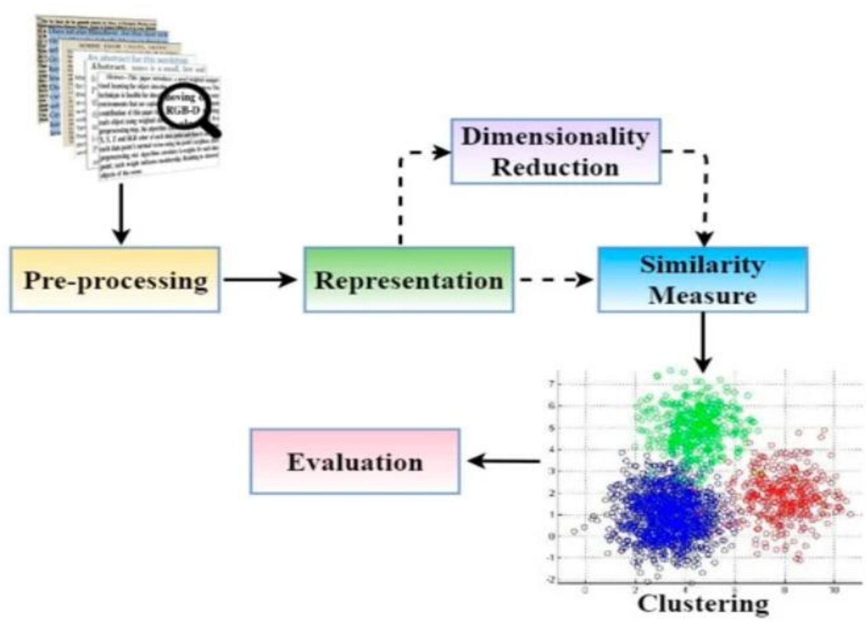


Figure 3.1 Block Diagram of Existing System

3.3 ALGORITHM AND METHODOLOGY

To achieve comprehensive client segmentation, the study employs three primary modules within its algorithm and methodology:

Data Collection:

The study collects data from the mall consumer Segmentation Dataset, including information on age, yearly income, consumption score, and other relevant consumer attributes.Data collection is conducted meticulously to ensure the accuracy and integrity of the dataset.

Clustering Analysis:

Utilizing the k-means, hierarchical clustering, and DBSCAN algorithms, the study conducts clustering analysis on the collected data.Each algorithm is applied to the dataset to segment clients based on their distinct characteristics, such as age, income, and consumption behavior.
The clustering results provide valuable insights into customer behavior patterns and preferences.

Evaluation and Optimization:

The clustering results are evaluated to determine the performance of each algorithm in segmenting clients effectively.Techniques for optimization may be employed to enhance the accuracy and reliability of the segmentation results.The study aims to identify the most suitable clustering algorithm for client segmentation based on the evaluation outcomes

3.4 SUMMARY

In summary, the study employs a multi-faceted approach to client segmentation using k-means, hierarchical clustering, and DBSCAN algorithms. By meticulously collecting and analyzing data from the mall consumer Segmentation Dataset, the study provides valuable insights that can inform strategic decision-making and improve market performance. Additionally, the study evaluates the performance of each clustering algorithm and aims to optimize the segmentation process for enhanced accuracy and reliability.While the existing system effectively employs clustering algorithms to gain insights into customer behavior patterns, it lacks sufficient privacy measures to safeguard sensitive consumer data adequately. Recognizing this limitation, the study advocates for the integration of privacy preservation techniques to enhance data security and protect consumer privacy. By incorporating privacy preservation measures alongside clustering analysis, the study ensures that sensitive information remains confidential throughout the segmentation process. This additional layer of privacy protection aligns with ethical considerations and regulatory requirements, reinforcing trust among stakeholders and mitigating potential risks associated with data breaches or unauthorized access. Overall, the integration of privacy preservation techniques enhances the integrity and reliability of the segmentation process while upholding consumer privacy rights and promoting responsible data management practices.

CHAPTER 4

PROPOSED SYSTEM

CHAPTER 4

PROPOSED SYSTEM

4.1 OVERVIEW

This section delineates the design and evaluation of an innovative privacy-preserving technique aimed at safeguarding sensitive data. The proposed methodology, illustrated in the above block diagram, integrates robust privacy mechanisms with effective data analysis techniques to ensure privacy while maintaining data utility and accuracy.

The proposed technique follows a structured workflow, divided into two main phases:

Phase 1: Privacy Preservation Phase This phase focuses on safeguarding individuals' privacy in datasets and comprises two modules:

(a) **Data Preprocessing Module:** In this module, raw data undergoes preprocessing steps to ensure data quality and consistency. This includes tasks such as data cleaning, normalization, and feature engineering.

(b) **Dimensionality Reduction Module:** Principal Component Analysis (PCA) is applied to reduce the dimensionality of the dataset while preserving its essential information. PCA-based feature selection aids in enhancing classification accuracy and reducing computational complexity.

Phase 2: Privacy-Preserving Clustering Phase This phase involves clustering perturbed datasets to extract meaningful insights while preserving privacy:

(a)**Differential Privacy Module:** Perturbed data undergoes further modification through the application of Differential Privacy mechanisms. These mechanisms ensure privacy guarantees while perturbing the datasets. The accuracy of the perturbed dataset is evaluated and compared with the original dataset to assess the impact of privacy preservation techniques.

(b) **Clustering and Evaluation Module:** The perturbed data is clustered using the K-means algorithm, adapted to preserve privacy. Traditional clustering techniques are adjusted to maintain privacy while extracting valuable insights from the data. This involves careful selection of perturbation methods, distance metrics, and privacy-preserving mechanisms to strike a balance between privacy protection and clustering accuracy.

Overall, the proposed technique aims to protect privacy while maintaining classification accuracy through a comprehensive framework encompassing data preprocessing, dimensionality reduction, clustering, and evaluation phases. By implementing PCA-based feature selection and Differential Privacy mechanisms, the technique ensures robust privacy preservation while enabling effective clustering and evaluation of perturbed datasets.The effectiveness of the approach is evaluated by comparing clustering accuracy before and after privacy preservation, followed by a comprehensive analysis and evaluation of the results to inform decision-making in real-world scenarios.

4.2 BLOCK DIAGRAM

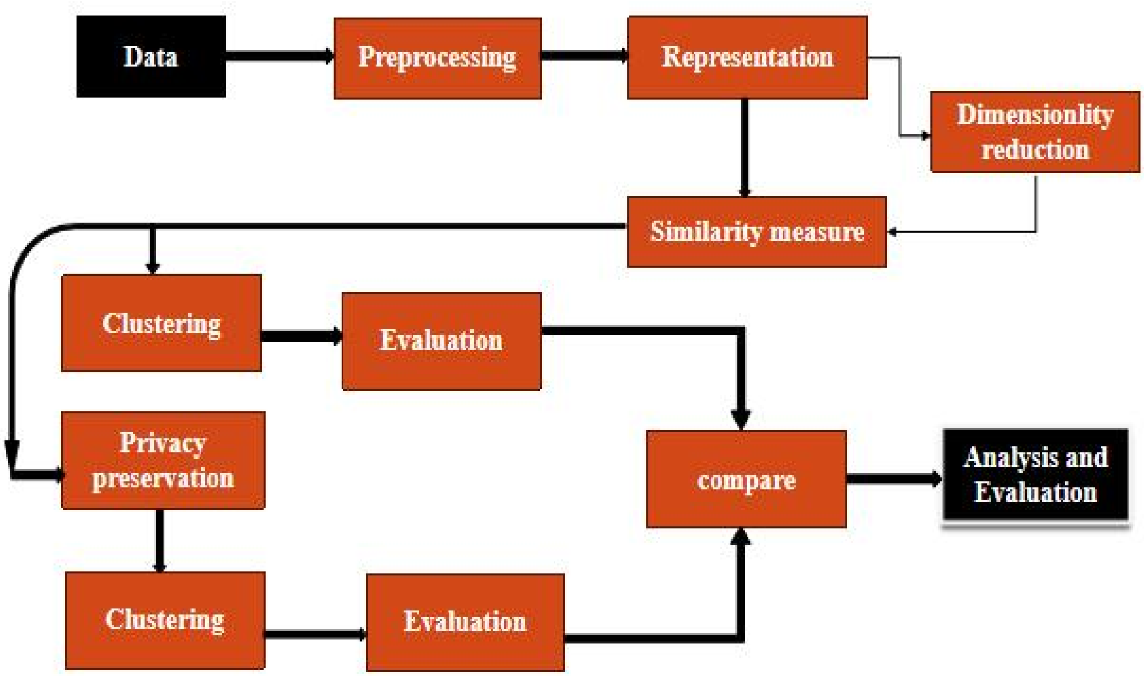


Figure 4.1 Block Diagram of Proposed System

4.3 ALGORITHM AND METHODOLOGY

4.3.1 Data Collection:

The Dataset Collection module involves acquiring the necessary data for analysis from various sources. This process may include accessing public datasets, collecting data through surveys or experiments, or obtaining data from external sources such as APIs or databases. The collected datasets should be relevant to the problem being addressed and should adhere to any legal or ethical considerations regarding data usage and privacy.

4.3.2 Preprocessing:

This module involves preparing the raw data for analysis by cleaning, transforming, and normalizing it. Common preprocessing techniques include handling missing values, outlier detection and removal, data normalization, and data scaling. The goal is to ensure that the data is in a suitable format for subsequent analysis steps.

4.3.3 Principal Component Analysis (PCA):

PCA is a dimensionality reduction technique used to identify patterns in high-dimensional data and represent it in a more compact form. It accomplishes this by transforming the original features into a new set of orthogonal variables called principal components. PCA helps in reducing the dimensionality of the data while preserving most of its variance, thus aiding in feature selection and simplifying subsequent analysis.

4.3.4 Clustering:

Differential privacy is a rigorous privacy framework that provides mathematical guarantees for protecting the privacy of individuals in datasets. In this module, Differential Privacy mechanisms are applied to perturb the data while ensuring that the privacy of individuals is preserved. This may involve adding noise to the data or applying other privacy-preserving transformations to prevent the disclosure of sensitive information about individuals.

4.3.5 Privacy Preservation:

Differential privacy is a rigorous privacy framework that provides mathematical guarantees for protecting the privacy of individuals in datasets. In this module, Differential Privacy mechanisms are applied to perturb the data while ensuring that the privacy of individuals is preserved. This may involve adding noise to the data or applying other privacy-preserving transformations to prevent the disclosure of sensitive information about individuals.

4.3.6 Clustering on Preserved Data:

Building upon the privacy preservation module, this module applies clustering algorithms, namely k-means, hierarchical, and spectral clustering, to the privacy-preserving data. By clustering the data after privacy preservation, the module assesses the impact of differential privacy on clustering accuracy and performance. This analysis provides insights into the trade-offs between privacy protection and clustering effectiveness.

4.3.7 Performance Evaluation:

The performance evaluation module employs metrics such as silhouette score, Davies-Bouldin index, and Calinski-Harabasz index to assess the quality of clustering results. These metrics evaluate the compactness, separation, and overall structure of the clusters generated by the clustering algorithms. By quantifying the performance of clustering techniques, this module facilitates objective comparisons and informs decision-making in the data analysis process.

4.4 SUMMARY

In summary, the study underscores the critical importance of understanding customer behavior patterns in the contemporary digital landscape for corporate success. Employing three distinct clustering algorithms - k-means, hierarchical clustering, and DBSCAN - the study aims to delve deeply into client segmentation. By meticulously analyzing factors such as age, yearly income, and consumption score, the study offers a comprehensive perspective on various consumer attributes using data from the mall consumer Segmentation Dataset. This insight serves as a valuable tool for adjusting marketing strategies, empowering stakeholders to make informed decisions and improve market performance. However, the existing system lacks sufficient privacy measures to safeguard sensitive consumer data adequately. Recognizing this limitation, the study advocates for the integration of privacy preservation techniques to enhance data security and protect consumer privacy. By incorporating privacy preservation measures alongside clustering analysis, the study ensures that sensitive information remains confidential throughout the segmentation process. This additional layer of privacy protection aligns with ethical considerations and regulatory requirements, reinforcing trust among stakeholders and mitigating potential risks associated with data breaches or unauthorized access. Overall, the integration of privacy preservation techniques enhances the integrity and reliability of the segmentation process while upholding consumer privacy rights and promoting responsible data management practices.

CHAPTER 5

IMPLEMENTATION SETUP

CHAPTER 5

IMPLEMENTATION SETUP

5.1 Environment Setup

Programming Environment:

The project was implemented using Python programming language due to its versatility and extensive libraries for data analysis and machine learning tasks. Python environments, managed using tools like Anaconda or virtualenv, were set up to ensure dependency management and reproducibility across different systems.

Library Installation:

Essential Python libraries such as NumPy, pandas, scikit-learn, and matplotlib were installed using package managers like pip or conda. These libraries provided functionalities for data manipulation, preprocessing, dimensionality reduction, clustering, and performance evaluation, streamlining the implementation process.

5.2 Data Collection

Data Sources and Integrity:

The breast cancer dataset and the wine quality dataset were obtained from reliable sources, ensuring their integrity and relevance to the research objectives. The provenance of the datasets was carefully documented to maintain transparency and reproducibility throughout the analysis process.

Dataset Description:

The breast cancer dataset comprises features related to tumor characteristics, including attributes such as tumor size, malignancy, and histological type. On the other hand, the wine quality dataset includes attributes related to wine composition, such as alcohol content, acidity levels, and quality ratings provided by experts or consumers.

5.3 Data Preprocessing Implementation

Data Cleaning:

Custom Python scripts were developed to address missing values, inconsistencies, and errors in the datasets. Techniques such as mean, median, or regression imputation for numerical attributes and mode imputation for categorical attributes were implemented using pandas DataFrame operations. Outliers and noise were detected and adjusted using statistical methods such as z-score or interquartile range (IQR).

Data Transformation:

Categorical variables were encoded into numerical form using techniques like one-hot encoding or label encoding, implemented using scikit-learn's preprocessing module. Numerical attributes underwent transformations such as logarithmic or square root transformations to address skewness, and scaling was applied using techniques like MinMaxScaler or StandardScaler to ensure uniformity across attributes.

Data Representation:

The datasets were represented as pandas DataFrames, ensuring that all relevant information was captured and structured for further analysis. Feature names and data types were carefully documented to maintain transparency and facilitate downstream processing steps.

5.4 Dimensionality Reduction using PCA:

PCA was implemented using scikit-learn's PCA module to reduce the dimensionality of the datasets. The original 30 features were transformed into 10 principal components, retaining essential information while reducing computational complexity. The transformed datasets were then ready for clustering analysis.

5.5 Clustering Implementation

K-means, hierarchical, and spectral clustering algorithms were implemented using scikit-learn's clustering module. Custom Python scripts were developed to partition the preprocessed datasets into distinct clusters based on their similarities. The number of clusters and other hyperparameters were optimized through experimentation to improve clustering performance.

5.6 Privacy Preservation using Differential Privacy

Differential privacy techniques were implemented using custom Python scripts to protect sensitive information while ensuring individuals' data privacy during the clustering process. This involved introducing controlled noise or perturbations to the preprocessed datasets, with the level of noise carefully calibrated to maintain privacy guarantees.

5.7 Clustering on Preserved Data

The clustering algorithms were reapplied to the perturbed datasets to evaluate the impact of privacy-preserving measures on clustering accuracy. Performance metrics such as silhouette score, Davies-Bouldin index, and Calinski-Harabasz score were computed to assess whether the clustering results remained meaningful and informative despite the introduction of privacy safeguards.

5.8 Performance Evaluation

The performance of each clustering algorithm was evaluated using metrics such as silhouette score, Davies-Bouldin index, and Calinski-Harabasz score. Custom Python scripts were developed to compute these metrics based on the clustering results, providing insights into the quality and effectiveness of the clustering techniques.

5.9 SUMMARY

The project encompassed the comprehensive setup and execution of a Python-based environment, incorporating crucial libraries like NumPy, pandas, scikit-learn, and matplotlib. With a meticulous focus on data preprocessing, the breast cancer and wine quality datasets underwent rigorous cleaning, transformation, and representation procedures to ensure optimal data quality and alignment with analytical objectives. Leveraging Principal Component Analysis (PCA), the dimensionality of the datasets was efficiently reduced, paving the way for the application of diverse clustering algorithms including K-means, hierarchical, and spectral clustering. An integral aspect of the project was the implementation of sophisticated differential privacy techniques, strategically integrated to safeguard sensitive data during the clustering process. Finally, an in-depth performance evaluation phase ensued, utilizing a diverse array of metrics to meticulously assess the efficacy and robustness of the clustering methodologies deployed, thus culminating in a comprehensive and insightful analysis of the datasets.

CHAPTER 6

RESULT AND INFERENCES

CHAPTER 6
RESULT AND INFERENCES

6.1 PERFORMANCE EVALUATION

Table 6.1 Descripion of the evaluation metrics

Performance Metrics	Range	Description
Silhouette Score	-1 to 1	Measures how similar an object is to its own cluster compared to other clusters. Higher score indicates better clustering
Davies–Bouldin Index	0 to Infinity	Evaluates the separation between clusters. Lower values mean better clustering.
Calinski-Harabasz Index	0 to Infinity	Based on the ratio of between-cluster dispersion to within-cluster dispersion. Higher index suggests better clustering.

6.2 RESULTS

The PCA dimensionality reduction technique effectively reduced the feature space while preserving the essential information, facilitating more efficient clustering and analysis. However, the application of differential privacy led to a reduction in clustering accuracy compared to clustering on the original datasets. Despite this trade-off, the privacy-preserving measures ensured the confidentiality and integrity of sensitive data, aligning with ethical and regulatory considerations. Among the clustering algorithms, K-means exhibited the highest accuracy, followed by hierarchical clustering and spectral clustering. These findings underscored the importance of balancing privacy preservation with clustering accuracy to make informed decisions in data-driven applications.

Before Differential Privacy:

i. Original dataset with accurate information.
ii. May contain sensitive attributes posing privacy risks.
iii. High utility and accuracy but vulnerable to security threats.

After Differential Privacy:

i. Perturbed dataset with privacy-preserving modifications.
ii. Protects sensitive attributes, enhancing privacy.
iii. May exhibit reduced utility and accuracy due to perturbation.
iv. Alters data distribution while improving security.

Table 6.2 Performance of K-Means clustering on original and modified Breast Cancer Dataset

Performance metrics of K-Means clustering	Before applying differential privacy	After applying differential privacy
Silhouette Score	0.35774	0.23156
Davies–Bouldin Index	1.25669	1.7351
Calinski-Harabasz Index	288.0915	157.1866

Table 6.3 Performance of Hierarchical clustering on original and modified Breast Cancer Dataset

Performance metrics of Hierarchical clustering	Before applying differential privacy	After applying differential privacy
Silhouette Score	0.29599	0.2278
Davies–Bouldin Index	1.38045	1.8134
Calinski-Harabasz Index	244.0943	136.6525

Table 6.4 Performance of Spectral clustering on original and modified Breast Cancer Dataset

Performance metrics of K-Means clustering	Before applying differential privacy	After applying differential privacy
Silhouette Score	0.35145	0.22458
Davies–Bouldin Index	1.26858	1.73861
Calinski-Harabasz Index	283.427	151.68537

Table 6.5 Performance of K-Means clustering on original and modified wine quality dataset

Performance metrics of K-Means clustering	Before applying differential privacy	After applying differential privacy
Silhouette Score	0.1808	0.3209
Davies–Bouldin Index	1.4588	0.8891
Calinski-Harabasz Index	275.3568	616.2626

Table 6.6 Performance of Hierarchical clustering on original and modified wine quality dataset

Performance metrics of K-Means clustering	Before applying differential privacy	After applying differential privacy
Silhouette Score	0.1495	0.26694
Davies–Bouldin Index	1.5263	0.95014
Calinski-Harabasz Index	225.7710	497.7541

Table 6.7 Performance of Spectral clustering on original and modified wine quality dataset

Performance metrics of K-Means clustering	Before applying differential privacy	After applying differential privacy
Silhouette Score	0.05982	0.2968
Davies–Bouldin Index	1.47772	0.8831
Calinski-Harabasz Index	217.039	574.0605

6.3 SUMMARY

In this pivotal module of the project, a meticulous evaluation of algorithmic performance is conducted to gauge the effectiveness and suitability of each clustering technique employed. Through the systematic utilization of a variety of metrics, ranging from the Silhouette Score to the Davies-Bouldin Index and the Calinski-Harabasz Index, a comprehensive understanding of each algorithm's strengths and weaknesses is attained. By presenting these metrics in a structured table format, the module facilitates a comparative analysis, enabling stakeholders to make well-informed decisions regarding the most appropriate clustering approach for their specific use case. Furthermore, the module serves as a crucial checkpoint in the project's journey, offering validation and assurance of the chosen methodologies. By rigorously assessing the performance of each algorithm across multiple dimensions, including cluster compactness, separation, and overall cohesion, the module ensures that the clustering techniques employed align closely with the project's objectives and requirements. Ultimately, this module stands as a cornerstone of the project, offering a rigorous and transparent evaluation framework that underpins the credibility and reliability of the clustering results. Through its systematic approach to algorithmic assessment and performance evaluation, the module empowers stakeholders to make data-driven decisions with confidence, ensuring the successful execution and deployment of clustering techniques in real-world scenarios.

CHAPTER 7

CONCLUSION AND FUTURE WORK

CHAPTER 7

CONCLUSION AND FUTURE WORK

Conclusion:

Integrating privacy-preserving techniques is essential in safeguarding sensitive data during the clustering process. As organizations increasingly rely on data-driven insights, maintaining individual privacy rights becomes paramount. By incorporating privacy algorithms like differential privacy or federated learning into clustering workflows, organizations can mitigate the risk of data breaches and unauthorized access. However, the application of privacy-preserving methods often introduces noise or distortion to the data, impacting clustering accuracy. Thus, striking a balance between privacy preservation and clustering accuracy is crucial. Moreover, dimensionality reduction techniques such as Principal Component Analysis (PCA) play a significant role in enhancing privacy preservation. PCA allows for the anonymization of sensitive features while retaining essential information, thereby facilitating more effective clustering while protecting individual privacy. Ultimately, prioritizing robust privacy mechanisms enables informed decision-making, fosters trust with stakeholders, and ensures responsible data stewardship in the era of big data and advanced analytics.

Future Work:

Moving forward, there are several avenues for further exploration and enhancement in the realm of privacy-preserving clustering. Firstly, research efforts could focus on developing more advanced privacy algorithms that strike an optimal balance between privacy protection and clustering accuracy. Additionally, investigating the integration of differential privacy techniques with emerging clustering algorithms, such as deep learning-based approaches, could yield improved performance in terms of both privacy and clustering quality. Furthermore, exploring the application of privacy-preserving clustering techniques in specific domains such as healthcare or finance, where data privacy is particularly critical, could provide valuable insights and contribute to the development of domain-specific privacy solutions.

REFERENCES

1. Gupta, Manoj Kr, and Pravin Chandra. "A comparative study of clustering algorithms." *2019 6th international conference on computing for sustainable global development (INDIACom)*. IEEE, 2019.
2. Murtagh, Fionn, and Pedro Contreras. "Algorithms for hierarchical clustering: an overview." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 2.1 (2012): 86-97.
3. Ratra, Ritu, et al. "Big data privacy preservation using principal component analysis and random projection in healthcare." Mathematical Problems in Engineering 2022 (2022).
4. Aldeen, Yousra Abdul Alsahib S., Mazleena Salleh, and Mohammad Abdur Razzaque. "A comprehensive review on privacy preserving data mining." SpringerPlus 4 (2015): 1-36.
5. Alotaibi, Khaled, et al. "Non-linear dimensionality reduction for privacy-preserving data classification." 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing. IEEE, 2012.
6. Granato, Daniel, et al. "Use of principal component analysis (PCA) and hierarchical cluster analysis (HCA) for multivariate association between bioactive compounds and functional properties in foods: A critical perspective." Trends in Food Science & Technology 72 (2018): 83-90.
7. Jha, Somesh, Luis Kruger, and Patrick McDaniel. "Privacy preserving clustering." Computer Security–ESORICS 2005: 10th European Symposium on Research in Computer Security, Milan, Italy, September 12-14, 2005. Proceedings 10. Springer Berlin Heidelberg, 2005.
8. Malhi, Arnaz, and Robert X. Gao. "PCA-based feature selection scheme for machine defect classification." IEEE transactions on instrumentation and measurement 53.6 (2004): 1517-1525.
9. Ding, Shifei, Liwen Zhang, and Yu Zhang. "Research on spectral clustering algorithms and prospects." 2010 2nd International Conference on Computer Engineering and Technology. Vol. 6. IEEE, 2010.
10. Deep Clustering: Advances, Challenges, and Future Directions Ren, Y., Pu, J., Yang, Z., Xu, J., Li, G., Pu, X., ... & He, L. (2022). Deep clustering: A comprehensive survey. arXiv preprint arXiv:2210.04142.
11. Toshniwal, Durga. "Clustering techniques for streaming data-a survey." 2013 3rd IEEE international advance computing conference (IACC). IEEE, 2013.
12. Bouchachia, Abdelhamid. "Dynamic clustering." Evolving Systems 3.3 (2012): 133-134.
13. Liu, Yingze. "Customer Segmentation in User Behavior Analysis: A Comparative Study of

Clustering Algorithms." Highlights in Business, Economics and Management 21 (2023): 758-764.

14. Teslenko D, Sorokina A, Smelyakov K, et al. Comparative Analysis of the Applicability of Five Clustering Algorithms for Market Segmentation. 2023 IEEE Open Conference of Electrical, Electronic and Information Sciences, 2023: 1-6
15. Kremers, B.J., Citrin, J., Ho, A. and van de Plassche, K.L., 2023. Two-step clustering for data reduction combining DBSCAN and k-means clustering. Contributions to Plasma Physics, 63(5-6), p.e202200177.
16. Mrong, S. G., Islam, S., Akter, S., Mukta, S. A., & Rikta, S. A. (2023). Assessing Regional Disparities in Bangladesh: A Comparative Cluster Analysis of Health, Education, and Demographic Indicators across Districts. Asian Journal of Language, Literature and Culture Studies, 6(3), 325-335.
17. Seng, K. P., Ang, L. M., Ngharamike, E., & Peter, E. (2023). Ridesharing and crowdsourcing for smart cities: technologies, paradigms and use cases. IEEE Access, 11, 18038-18081.
18. Song, Youngho, Hyeong-Jin Kim, Hyun-Jo Lee, and Jae-Woo Chang. "A Parallel Privacy-Preserving k-Means Clustering Algorithm for Encrypted Databases in Cloud Computing." Applied Sciences 14, no. 2 (2024): 835.
19. Kushwah, S., Sharma, N., & Das, S. (2022). Feature-Based Consumer Healthcare Sentiments Analysis And Comparative Analysis Of Big Data Using Multiple Clustering Techniques. JOURNAL OF HARBIN INSTITUTE OF TECHNOLOGY, 54(5), 2022.
20. Kremers, B. J., Citrin, J., Ho, A., & van de Plassche, K. L. (2023). Two-step clustering for data reduction combining DBSCAN and k-means clustering. Contributions to Plasma Physics, 63(5-6), e202200177.
21. Liang, James, Tianfei Zhou, Dongfang Liu, and Wenguan Wang. "Clustseg: Clustering for universal segmentation." arXiv preprint arXiv:2305.02187 (2023).
22. Majeed, A., & Lee, S. (2020). Anonymization techniques for privacy preserving data publishing: A comprehensive survey. IEEE access, 9, 8512-8545.
23. Ram Mohan Rao, P., Murali Krishna, S. and Siva Kumar, A.P., 2018. Privacy preservation techniques in big data analytics: a survey. Journal of Big Data, 5(1), p.33

APPENDIX A

SOURCE CODE

APPENDIX A
SOURCE CODE

PRINCIPAL COMPONENT ANALYSIS:

```
#importing libraries:
import pandas as pd
from sklearn.datasets import load_breast_cancer
from sklearn.metrics import silhouette_score, davies_bouldin_score, calinski_harabasz_score

data = load_breast_cancer()
df = pd.DataFrame(data.data, columns=data.feature_names)
df['label'] = data.target

from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()
scaled_features = scaler.fit_transform(df.drop('label', axis=1))

from sklearn.decomposition import PCA
n_components = 10

pca = PCA(n_components=n_components)
pca_features = pca.fit_transform(scaled_features)
```

DIFFERENTIAL PRIVACY ALGORITHM:

```
import numpy as np
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
# Apply differential privacy
def add_noise(data, epsilon):
    # Add Laplace noise to the data
    noise = np.random.laplace(scale=1/epsilon, size=data.shape)
```

```
    return data + noise

epsilon = 1.0  # Privacy parameter (you can adjust this value)
noisy_data = add_noise(pca_features, epsilon=epsilon)

def mean_squared_error(original_data, noisy_data):
    # Calculate Mean Squared Error (MSE)
    mse = ((original_data - noisy_data) ** 2).mean()
    return mse

# Calculate MSE
mse = mean_squared_error(pca_features, noisy_data)
print("Mean Squared Error (MSE) between original and perturbed data:", mse)

#accuracy score of privacy preserved data
def mean_absolute_error(original_data, noisy_data):
    # Calculate Mean Absolute Error (MAE)
    mae = np.abs(original_data - noisy_data).mean()
    return mae

def custom_accuracy(original_data, noisy_data):
    # Binarize the data (0 if original value <= 0, 1 otherwise)
    binarized_original = np.where(original_data <= 0, 0, 1)
    binarized_noisy = np.where(noisy_data <= 0, 0, 1)

    # Calculate Accuracy Score
    acc_score = accuracy_score(binarized_original, binarized_noisy)
    return acc_score

# Calculate MAE and Accuracy Score
mae = mean_absolute_error(pca_features, noisy_data)
acc_score = custom_accuracy(pca_features, noisy_data)
```

```
print("Mean Absolute Error (MAE) between original and perturbed data:", mae)
print("Accuracy Score between original and perturbed data:", acc_score)
```

CLUSTERING ON PRIVACY PRESERVED DATASET:

K-Means

```
import matplotlib.pyplot as plt
# Function to calculate WCSS
def calculate_wcss(data, max_clusters):
    wcss = []
    for i in range(1, max_clusters + 1):
        kmeans = KMeans(n_clusters=i, init='k-means++', random_state=42)
        kmeans.fit(data)
        wcss.append(kmeans.inertia_)
    return wcss

# Plot WCSS
max_clusters = 10  # Maximum number of clusters to try
wcss_values = calculate_wcss(noisy_data, max_clusters)

plt.plot(range(1, max_clusters + 1), wcss_values, marker='o')
plt.title('Elbow Method for Optimal K')
plt.xlabel('Number of Clusters')
plt.ylabel('WCSS')
plt.xticks(range(1, max_clusters + 1))
plt.grid(True)
plt.show()
# Clustering
def kmeans_clustering(X, n_clusters):
    kmeans = KMeans(n_clusters=n_clusters, random_state=42)
    cluster_labels = kmeans.fit_predict(X)
```

```
    return cluster_labels

n_clusters = 2  # Number of clusters
kmeans_labels = kmeans_clustering(noisy_data, n_clusters)

# Plot clustered data
plt.scatter(noisy_data[:, 0], noisy_data[:, 1], c=kmeans_labels, cmap='viridis', alpha=0.5)
plt.xlabel('Principal Component 1')
plt.ylabel('Principal Component 2')
plt.title('K-means Clustering with Noisy Data')
plt.colorbar(label='Cluster')
plt.show()

clustering_metrics = evaluate_clustering(noisy_data, kmeans_labels)
print("Clustering Metrics:")
print("Silhouette Score:", clustering_metrics[0])
print("Davies-Bouldin Score:", clustering_metrics[1])
print("Calinski-Harabasz Score:", clustering_metrics[2])
```

Hierarchical Clustering

```
from sklearn.cluster import AgglomerativeClustering
import matplotlib.pyplot as plt

# Hierarchical clustering
def hierarchical_clustering(X, n_clusters):
    # Perform hierarchical clustering
    hc = AgglomerativeClustering(n_clusters=n_clusters, linkage='ward')
    cluster_labels = hc.fit_predict(X)

    return cluster_labels
```

```
# Perform hierarchical clustering
n_clusters = 2  # Number of clusters
hierarchical_labels = hierarchical_clustering(noisy_data, n_clusters)

# Plot clustered data
plt.scatter(noisy_data[:, 0], noisy_data[:, 1], c=hierarchical_labels, cmap='viridis', alpha=0.5)
plt.xlabel('Principal Component 1')
plt.ylabel('Principal Component 2')
plt.title('Hierarchical Clustering with Noisy Data')
plt.colorbar(label='Cluster')
plt.show()

from sklearn.metrics import silhouette_score, davies_bouldin_score, calinski_harabasz_score
import scipy.cluster.hierarchy as sch

# Hierarchical clustering with dendrogram
def hierarchical_clustering_with_dendrogram(X, n_clusters):
    # Perform hierarchical clustering
    hc = AgglomerativeClustering(n_clusters=n_clusters, linkage='ward')
    cluster_labels = hc.fit_predict(X)

    # Plot dendrogram
    plt.figure(figsize=(10, 6))
    dendrogram = sch.dendrogram(sch.linkage(X, method='ward'))
    plt.title('Dendrogram')
    plt.xlabel('Samples')
    plt.ylabel('Distance')
    plt.show()

    # Performance metrics
    silhouette = silhouette_score(X, cluster_labels)
    db_score = davies_bouldin_score(X, cluster_labels)
```

```
    calinski_score = calinski_harabasz_score(X, cluster_labels)

    return cluster_labels, silhouette, db_score, calinski_score

# Perform hierarchical clustering with dendrogram
n_clusters = 2  # Number of clusters
hierarchical_labels,silhouette,db_score,calinski_score =
hierarchical_clustering_with_dendrogram(noisy_data, n_clusters)

# Print performance metrics
print("Performance Metrics:")
print("Silhouette Score:", silhouette)
print("Davies-Bouldin Score:", db_score)
print("Calinski-Harabasz Score:", calinski_score)
```

Spectral Clustering:

```
from sklearn.cluster import SpectralClustering
from sklearn.metrics import silhouette_score, davies_bouldin_score, calinski_harabasz_score
import matplotlib.pyplot as plt

# Spectral clustering without cluster naming
def spectral_clustering(X, n_clusters):
    # Perform spectral clustering
    sc = SpectralClustering(n_clusters=n_clusters, affinity='nearest_neighbors', random_state=42)
    cluster_labels = sc.fit_predict(X)

    # Performance metrics
    silhouette = silhouette_score(X, cluster_labels)
    db_score = davies_bouldin_score(X, cluster_labels)
    calinski_score = calinski_harabasz_score(X, cluster_labels)

    return cluster_labels, silhouette, db_score, calinski_score
```

```
# Perform spectral clustering without cluster naming
n_clusters = 2
spectral_labels, silhouette, db_score, calinski_score = spectral_clustering(noisy_data, n_clusters)

# Plot clustered data with cluster labels
for label in np.unique(spectral_labels):
    plt.scatter(noisy_data[spectral_labels == label, 0], noisy_data[spectral_labels == label, 1], label=f'Cluster {label}')

# Add cluster labels to the plot
plt.text(noisy_data[spectral_labels == 0, 0].mean(), noisy_data[spectral_labels == 0, 1].mean(), '0', horizontalalignment='center', verticalalignment='center', fontsize=12)
plt.text(noisy_data[spectral_labels == 1, 0].mean(), noisy_data[spectral_labels == 1, 1].mean(), '1', horizontalalignment='center', verticalalignment='center', fontsize=12)

plt.xlabel('Principal Component 1')
plt.ylabel('Principal Component 2')
plt.title('Spectral Clustering with Noisy Data')
plt.legend()
plt.show()

# Print performance metrics
print("Performance Metrics:")
print("Silhouette Score:", silhouette)
print("Davies-Bouldin Score:", db_score)
print("Calinski-Harabasz Score:", calinski_score)
```

APPENDIX B

SNAPSHOTS

APPENDIX B

SNAPSHOTS

B.1 Breast Cancer Dataset

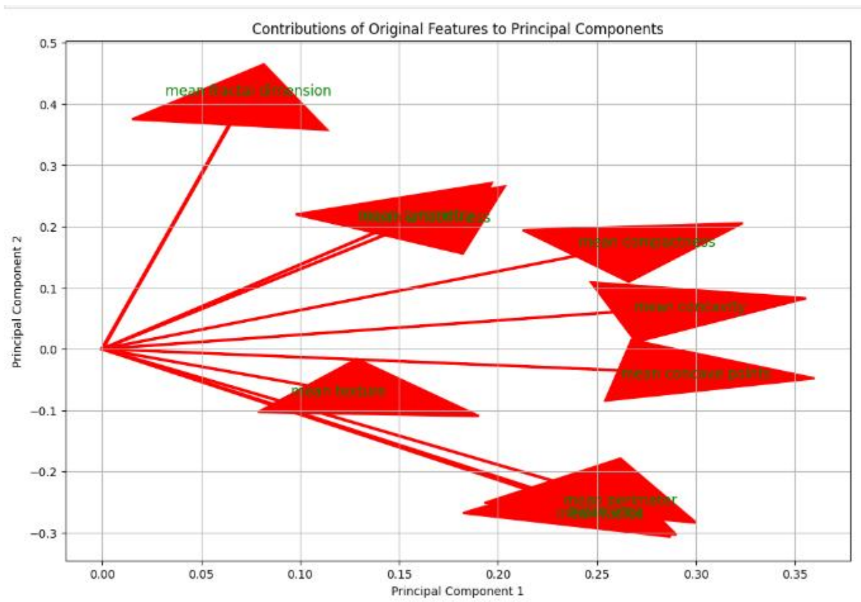


Figure B.1 Contributions of original features to principal components of breast cancer dataset

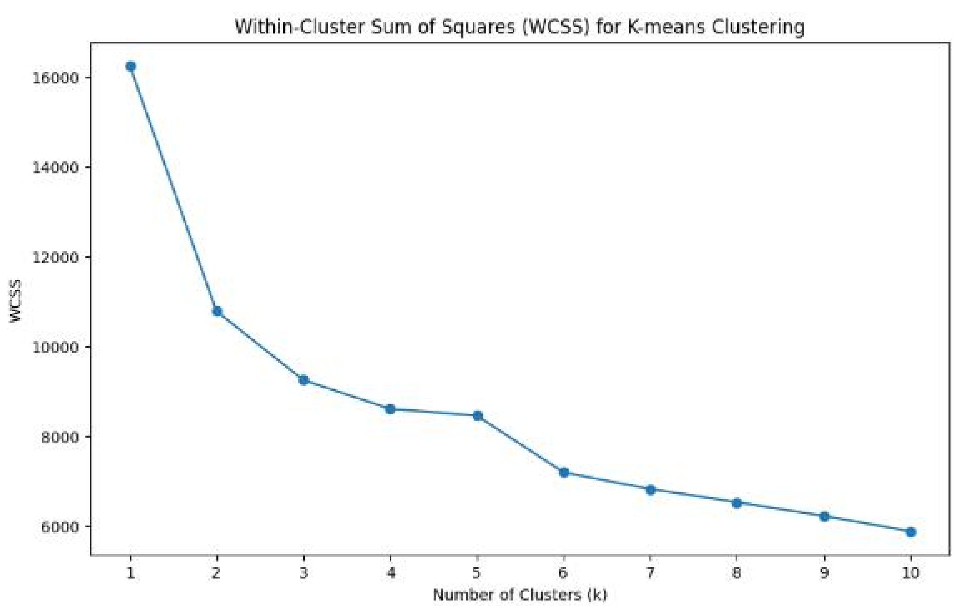


Figure B.2 Elbow method to compute optimal number of clusters in original breast cancer dataset

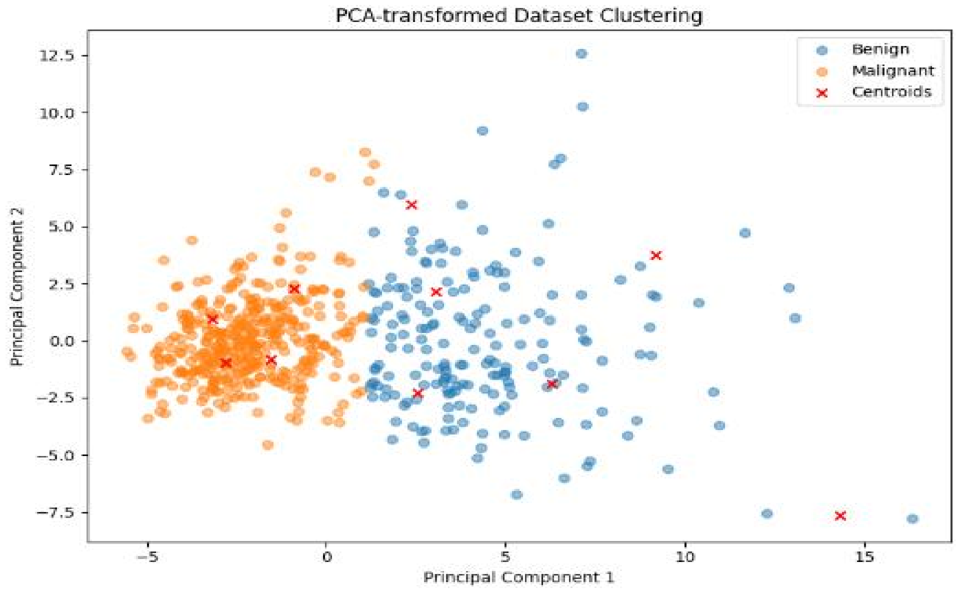


Figure B.3 K-Means clustering on original breast cancer dataset

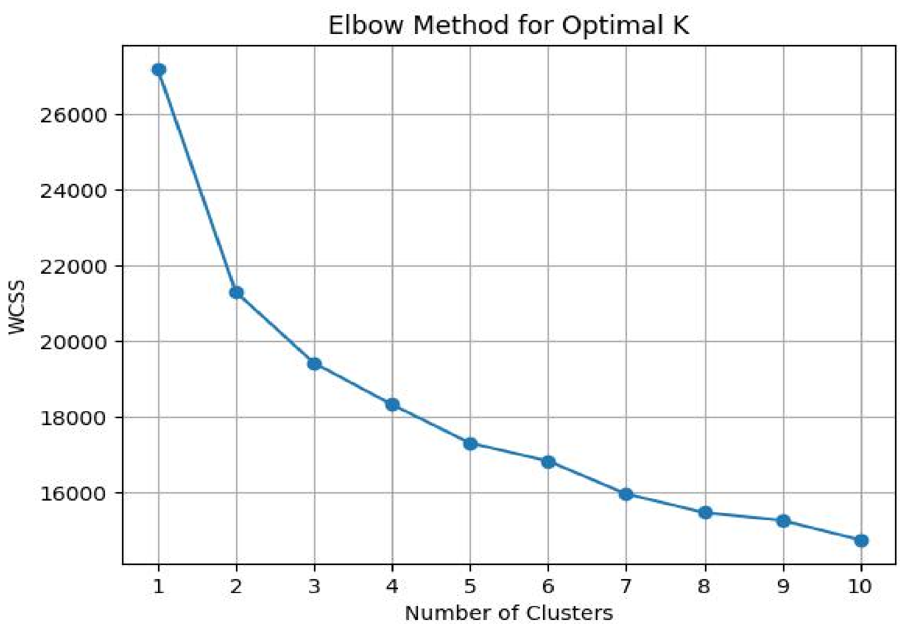


Figure B.4 Elbow method to compute optimal number of clusters in preserved breast cancer dataset

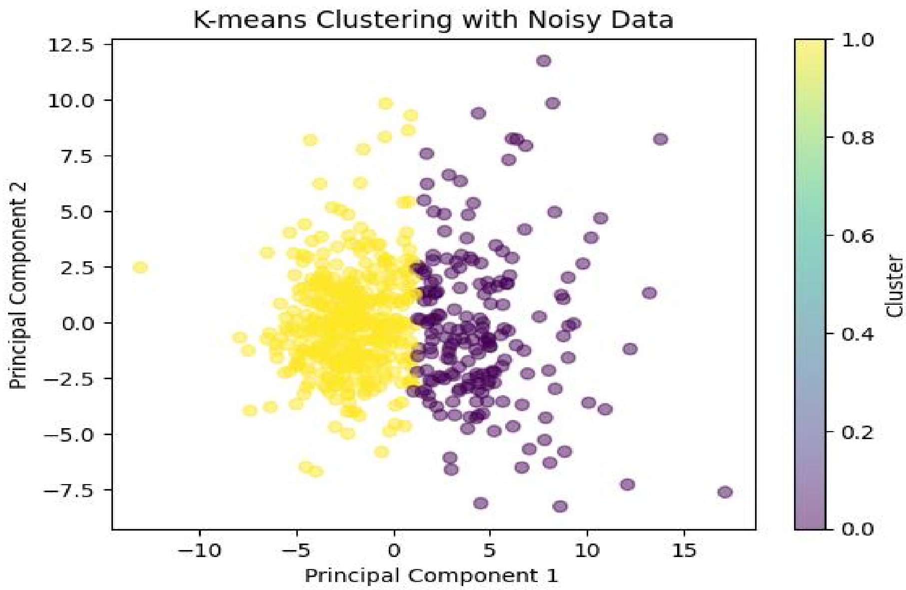


Figure B.5 K-Means clustering on Preserved breast cancer dataset

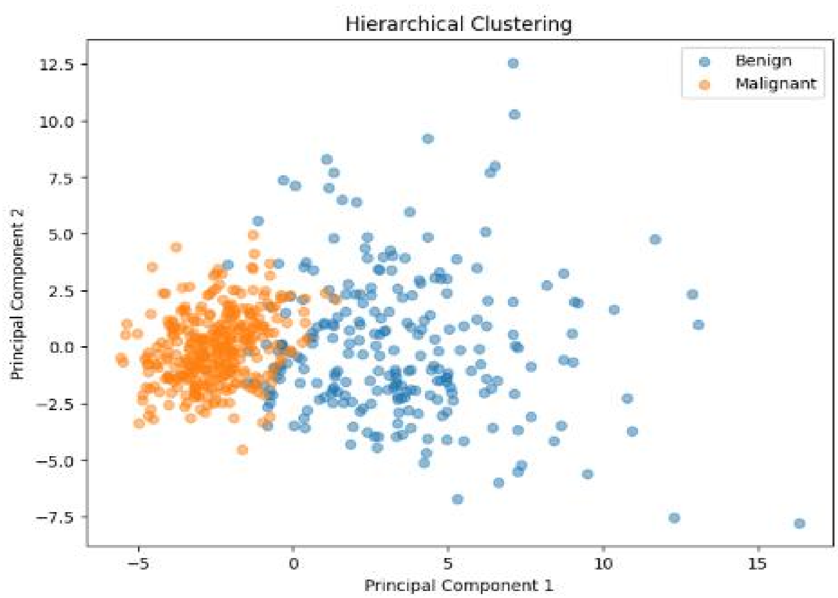


Figure B.6 Hierarchical clustering on original breast cancer dataset

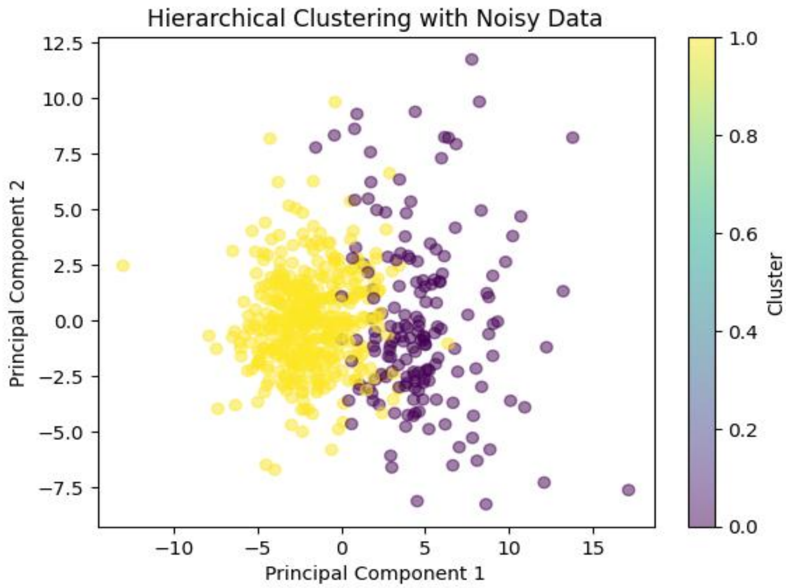


Figure B.7 Hierarchical clustering on preserved breast cancer dataset

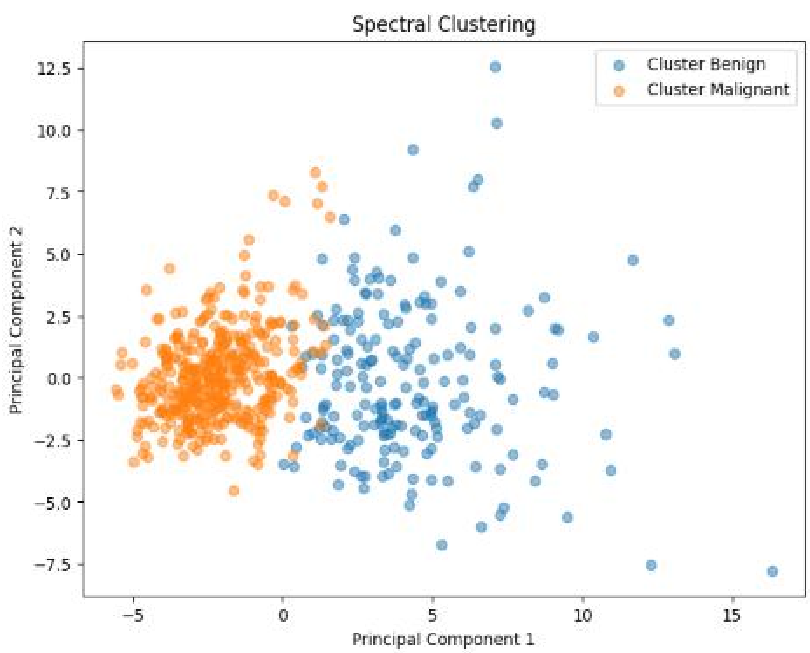


Figure B.8 Spectral clustering on original breast cancer dataset

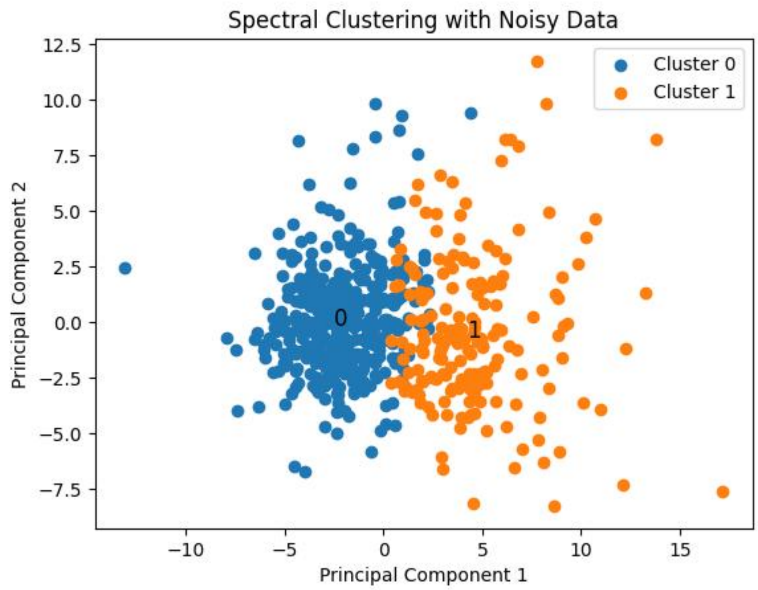


Figure B.9 Spectral clustering on preserved breast cancer dataset

B.2 Wine Quality Dataset

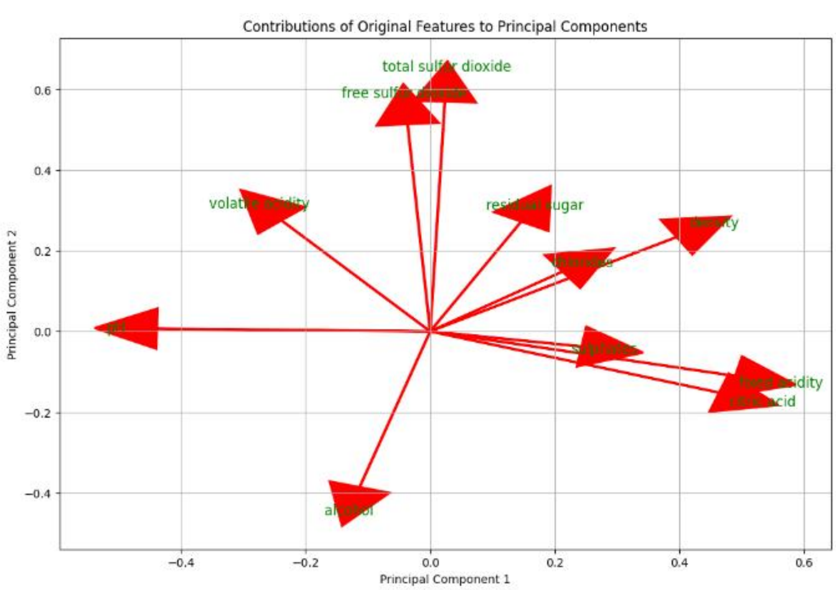


Figure B.10 Contributions of original features to principal components of wine quality dataset

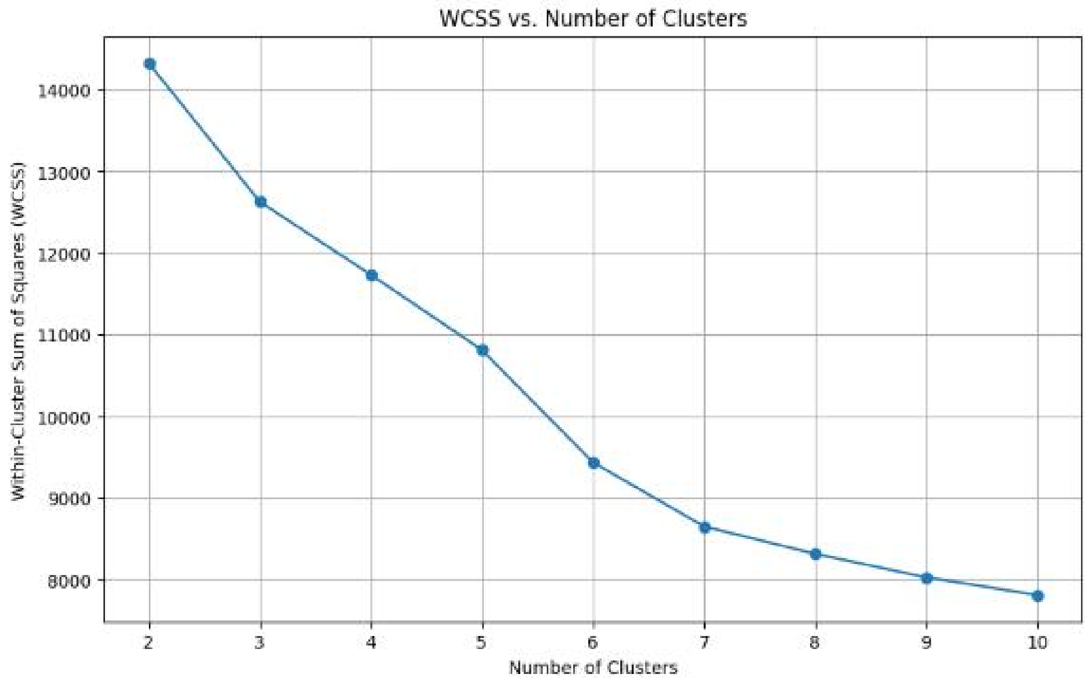


Figure B.11 Elbow method to compute optimal number of clusters in original wine quality dataset

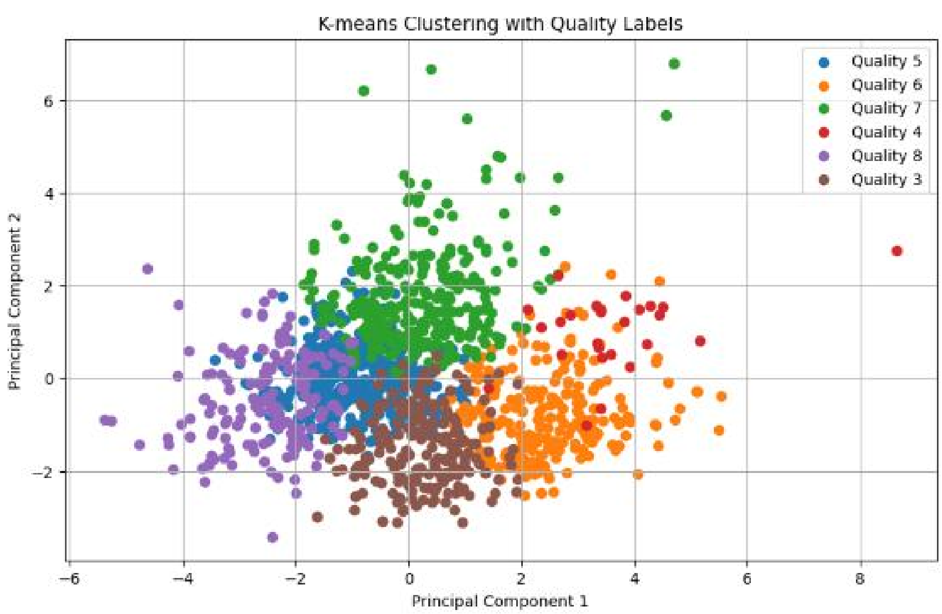


Figure B.12 K-Means clustering on original wine quality dataset

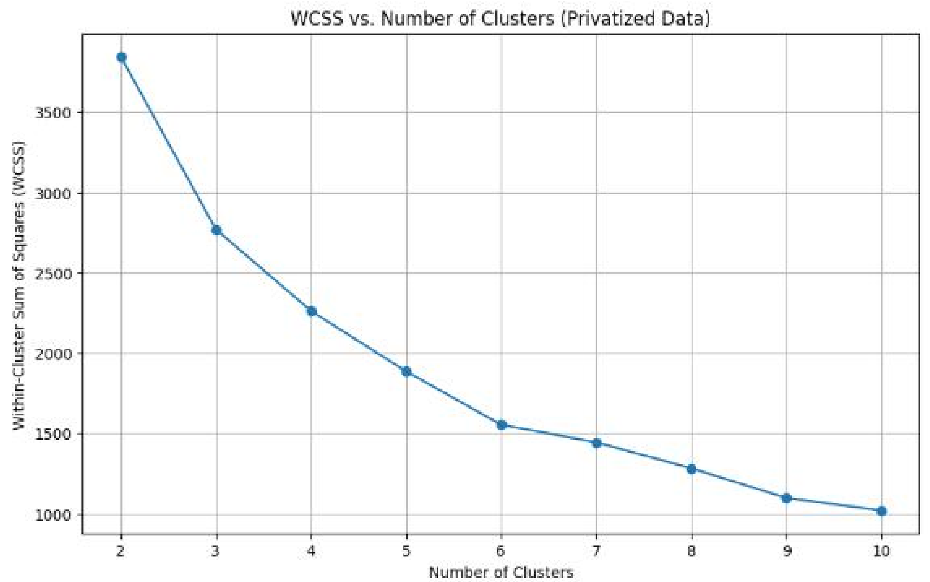


Figure B.13 Elbow method to compute optimal number of clusters in preserved wine quality dataset

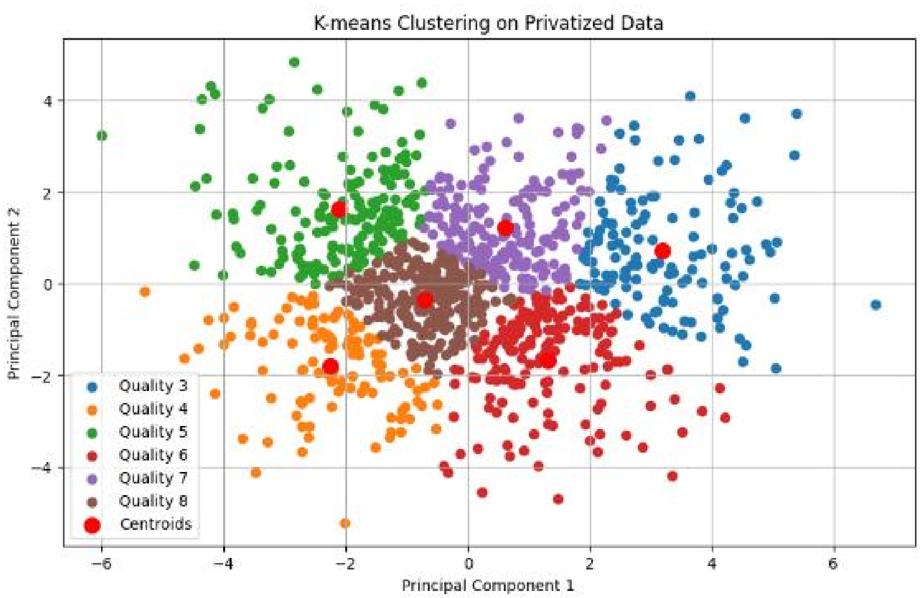


Figure B.14 K-Means clustering on Preserved wine quality dataset

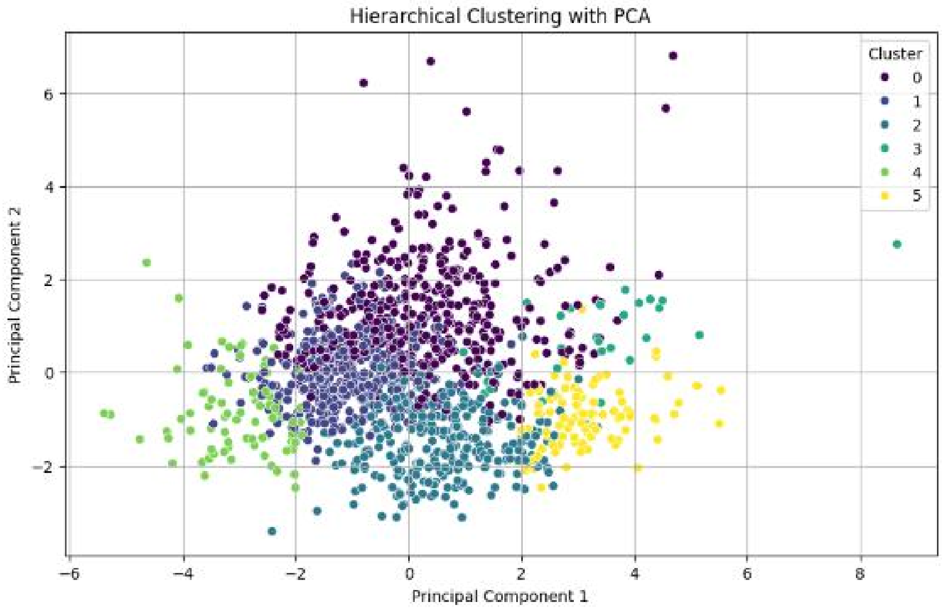


Figure B.15 Hierarchical clustering on original wine quality dataset

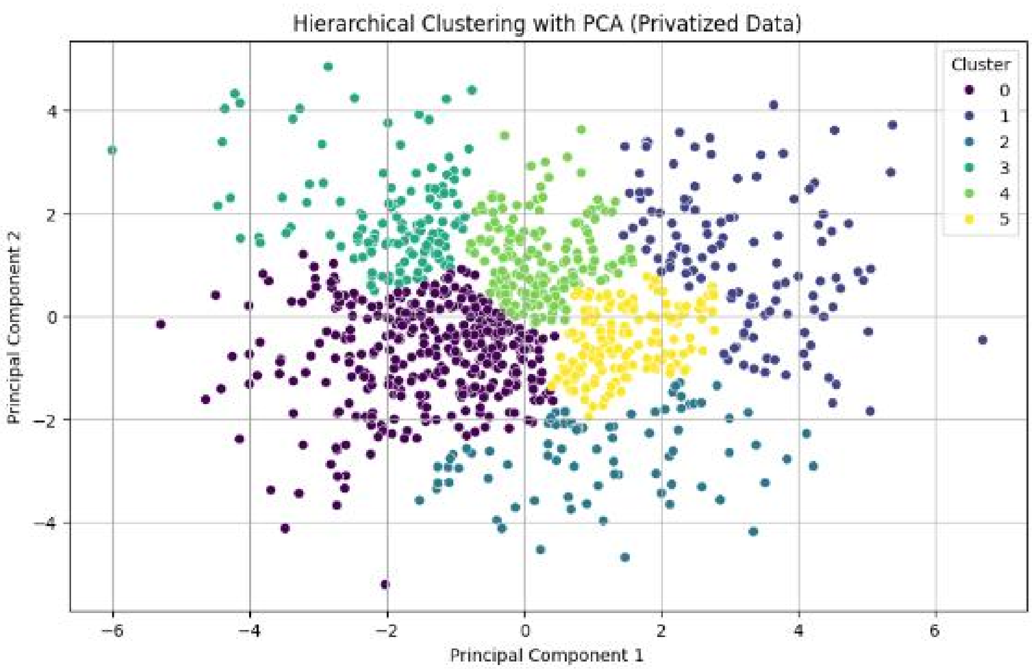


Figure B.16 Hierarchical clustering on Preserved wine quality dataset

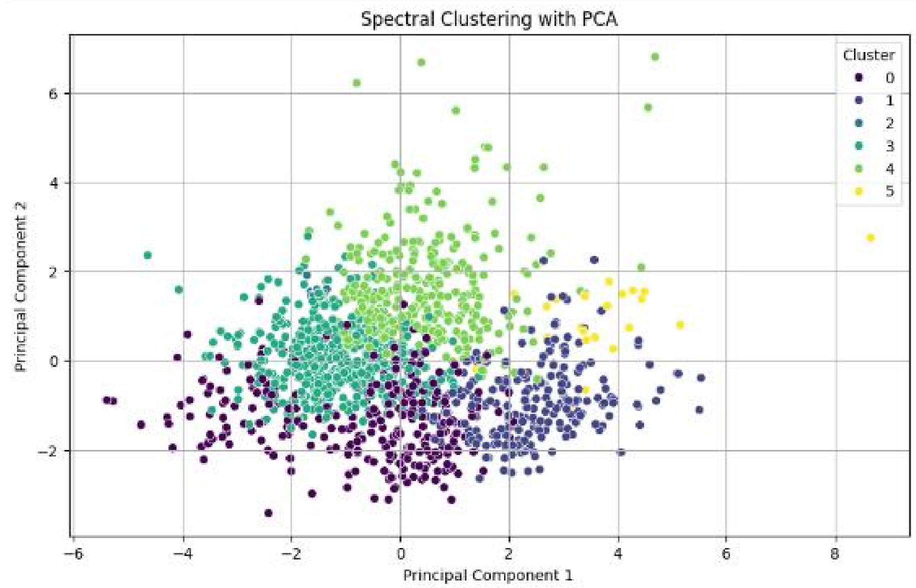


Figure B.17 Spectral clustering on original wine quality dataset

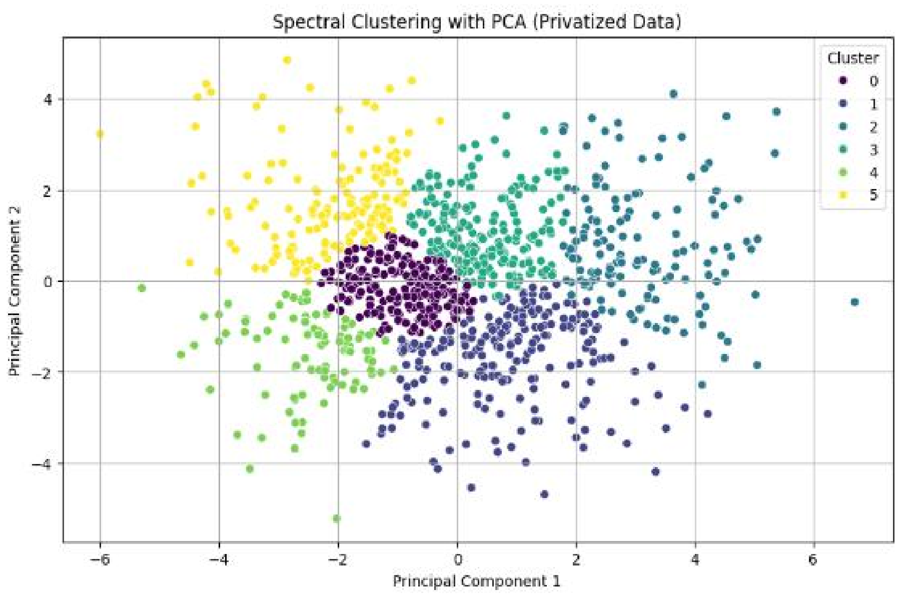


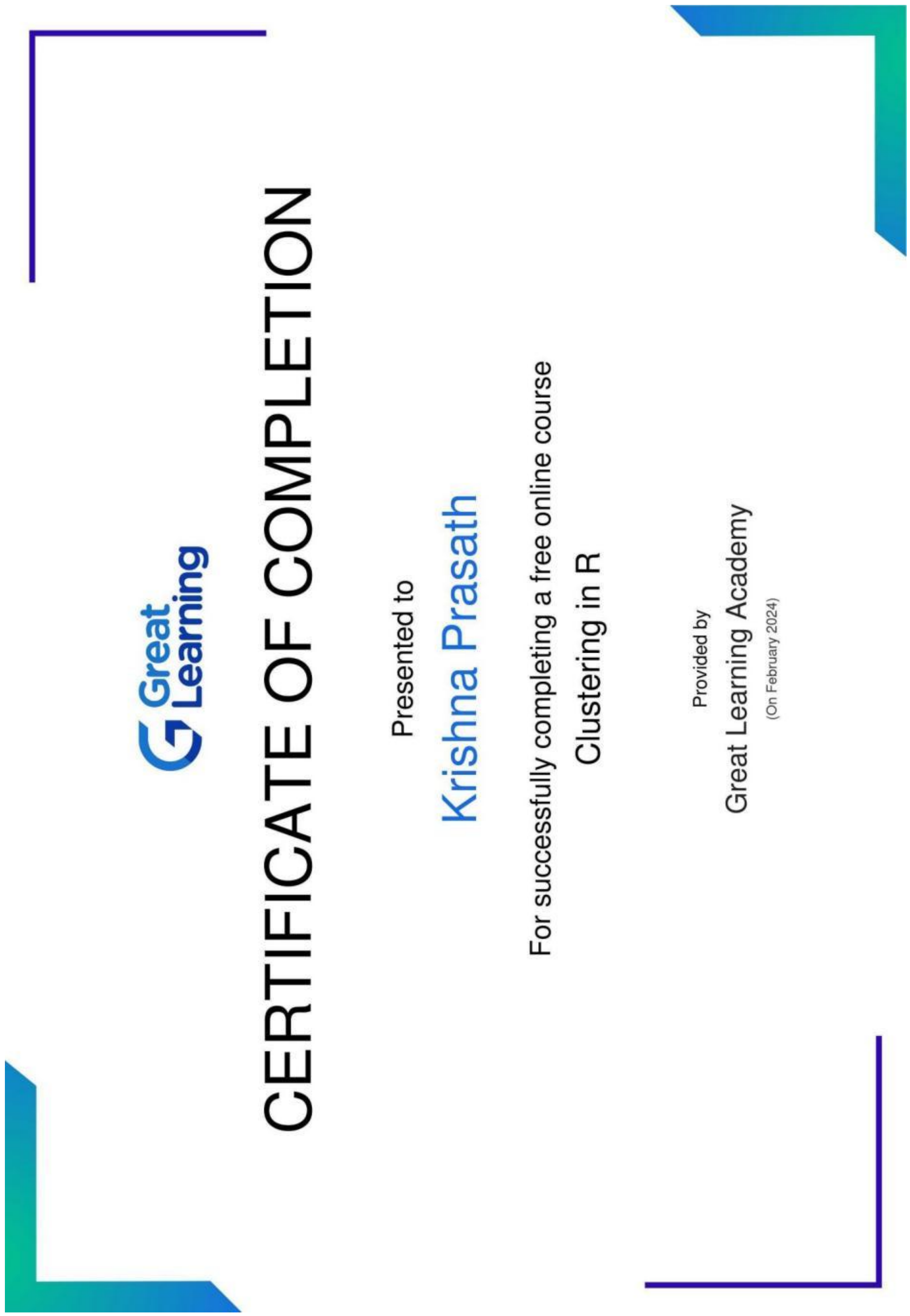
Figure B.18 Spectral clustering on Preserved wine quality dataset

APPENDIX C

CERTIFICATES

APPENDIX C

CERTIFICATES



Google

COURSE
CERTIFICATE

Feb 9, 2024

HARINI P

has successfully completed

Crash Course on Python

an online non-credit course authorized by Google and offered through Coursera

Google

Google

Verify at:
https://coursera.org/verify/MZCA3WPKSE6L
Coursera has confirmed the identity of this individual and their participation in the course.

simplilearn | SkillUP

CERTIFICATE OF COMPLETION

Sanofer Niswan S

has successfully completed the online course:

Introduction to Machine Learning with R

This professional has demonstrated initiative and a commitment to deepening their skills and advancing their career. Well done!

03rd Feb 2024

Certificate code : 4845429

Krishna Kumar
CEO, Simplilearn

VERIFIED