TECHNICAL REPORT

ADVANCED COMPUTING AND BIG DATA

CLUSTERING AND DATA VISUALISATION USING BREAST CANCER DATASET

OLEH

INDAH FITRIALITA

2206130744

DOSEN PENGAMPU

RISMAN ADNAN, PH.D

PROGRAM STUDI MAGISTER MATEMATIKA

UNIVERSITAS INDONESIA

DEPOK

2023

DAFTAR ISI

BAB I

INTRODUCTION

The increasing amount of data generated in various fields and industries has led to the development of techniques that help in understanding and analyzing many data. Clustering and data visualization are two such techniques that play a vital role in making sense of complex dan big data.

Breast cancer is one of the most common type of cancer among women. With the help of various screening and diagnostic tests, detection of breast cancer can greatly improve survival rates. However, the analysis and interpretation of large of data generated by these tests can be challenging. Clustering and data visualization techniques can help us identify patterns and relationships in the data, which can aid in the development of effective diagnostic and treatment strategies.

In this technical report, with ChatGPT help will show the utilization of clustering and data visualization techniques by analyzing the breast cancer dataset obtained from Kaggle. The analysis will commence with the exploration of the dataset and identification of pertinent features for analysis. Following that, we will utilize three clustering algorithms such as K-Means, Hierarchical, and DBSCAN (Density-Based Spatial Clustering of Applications with Noise) clustering to categorize the data into relevant clusters based on the identified features. Lastly, we will employ visualization methods such as scatter plots and heat maps to present the clustered data in a manner that is easy to interpret.

The aim of this report is to provide a comprehensive overview of clustering and data visualization techniques and demonstrate their usefulness in analyzing breast cancer data. The results of this analysis can have significant implications for the development of better diagnostic and treatment strategies for breast cancer.

BAB II

LITERATURE REVIEW

In this chapters, theories related to the thecnical report topic are presented, namely clustering and data visualisation.

2.1 Clustering

Clustering is a type of unsupervised learning thecnique in data mining that involves the grouping of similar dataset into group or cluster. The purpose of clustering is to indentify patterns, structures and relationship every data thtat are not easily visible using other analytical techniques. The process of clustering involves assigning each data point into a cluster based on their similarities or distance from other data point.

Here are clustering models or algorithms that can be used to perform clustering, such as :

1 K-Means Clustering

This algorithm divides a dataset into k cluster based on the distance between the data point and their respective cluster centers. The purpose is to minimize the sum of squared distances betweet each data point and its assigned cluster center.

The objective function is known as the Within-Cluster Sum of Squares (WCSS) or the Inertia. The WCSS measures the compactness of the clusters, with lower values indicating more tightly packed clusters. Optimal number of clusters k can be determined by using elbow method or silhouette score.

2 Hierarchical Clustering

Hierarchical clustering model involves grouping data points into a tree-like structure called a dendrogram that constructed by iteratively merging the most similar clusters untill all data points are in a single cluster. It does not require the specification of the number of clusters beforehand . There are two types of Hierarchical Clustering, such as Agglomerative Clustering and Divisive Clustering.

3 Density-Based Spatial Clustering of Applications with Noise

2.2 Data Visualisation

BAB III

RESULT

BAB IV

CONCLUSION