# BAB I
# PENDAHULUAN

## 1.1. Latar Belakang

Algoritma klastering adalah teknik penting dalam *machine learning* yang digunakan untuk mengelompokkan data berdasarkan kemiripan atau kedekatan tertentu. Salah satu algoritma klastering yang populer adalah *K-Medoids*. Algoritma ini mirip dengan K-Means, namun lebih tahan terhadap outlier karena menggunakan medoids (data yang berada di pusat klaster) sebagai pusat klaster, bukan *centroid* yang merupakan rata-rata dari seluruh data dalam klaster (Nurdin et al., 2024). Namun, pada dataset berukuran besar, algoritma *K-Medoids* sering mengalami tantangan dalam hal prediktabilitas dan stabilitas jumlah iterasi yang dibutuhkan untuk mencapai hasil akhir yang optimal (Nurdin et al., 2024).

*K-Medoids* adalah algoritma klastering yang mengelompokkan data dengan memilih data yang paling representatif sebagai pusat kluster, sehingga lebih stabil dan tidak terpengaruh oleh nilai ekstrem (Almazar et al., 2024). Proses ini melibatkan pemilihan medoids dari dataset, yang berfungsi sebagai representasi pusat kluster, sehingga meningkatkan ketahanan algoritma terhadap outlier dan memberikan hasil yang lebih robust dibandingkan dengan metode lain seperti K-Means. Pemilihan medoids yang tepat menjadi kunci dalam menghasilkan kluster yang akurat dan informatif, terutama dalam konteks dataset yang besar dan kompleks (Fajriana, 2021).

Salah satu aspek kritis dalam kinerja algoritma *K-Medoids* adalah inisiasi medoids awal. Inisiasi yang buruk dapat menyebabkan algoritma membutuhkan lebih banyak iterasi dan menghasilkan klaster yang kurang optimal. Oleh karena itu, diperlukan metode yang efektif untuk inisiasi medoids agar dapat meningkatkan kinerja algoritma *K-Medoids*, baik dari segi jumlah iterasi maupun kualitas klaster yang dihasilkan (Miraftabzadeh et al., 2013).

Metode *Z-Score*, di sisi lain, merupakan teknik statistik yang digunakan untuk menormalkan data dengan cara mengubah setiap nilai menjadi skor yang

menunjukkan seberapa jauh nilai tersebut dari rata-rata dalam satuan deviasi standard (Schubert, at al., 2022). Menggunakan *Z-Score*, data yang memiliki skala yang berbeda dapat dibandingkan secara langsung, dan ini memungkinkan pemilihan medoids awal yang lebih representatif. Metode ini tidak hanya meningkatkan akurasi penentuan pusat *cluster* tetapi juga membantu dalam mempercepat konvergensi algoritma *cluster* (Henderi, et al., 2021).

Penelitian terdahulu menunjukkan berbagai aplikasi optimasi algoritma *K-Medoids clustering*. Cynthia et al. (2020) membahas pentingnya pengoptimalan algoritma k-medoids dalam segmentasi pelanggan. Penelitian ini mengusulkan penggunaan algoritma *Particle Swarm Optimization* (PSO) untuk mengoptimalkan pemilihan pusat cluster awal. Firzada, et al., (2021) menggunakan *K-Medoids* untuk untuk mengklasifikasikan mahasiswa berdasarkan ketepatan waktu dalam menyelesaikan masa studi mereka. Algoritma k-medoids diterapkan untuk mengelompokkan mahasiswa yang menyelesaikan studi tepat waktu dan yang tidak. Sementara itu, Dinata, R.K., et al (2021) mengusulkan optimasi algoritma *K-Medoids* dengan menggunakan algoritma *purity*, yang berfokus pada upaya mengurangi jumlah iterasi yang dibutuhkan oleh algoritma k-medoids dalam proses clustering. Hasil penelitian tersebut menunjukkan bahwa algoritma *purity k-medoids* menghasilkan validitas cluster yang lebih baik dibandingkan algoritma *k-medoids* konvensional.

Berdasarkan penelitian terdahulu, dalam penelitian ini penulis mengusulkan penggunaan metode *Z-Score* untuk inisiasi medoids pada algoritma *K-Medoids.* Metode *Z-Score* dikenal karena kemampuannya dalam menormalkan data dan mengidentifikasi titik pusat yang lebih representatif dalam dataset. Dengan menerapkan metode *Z-Score*, diharapkan medoids awal yang dipilih lebih mendekati pusat sebenarnya dari klaster, sehingga dapat meminimalisasi jumlah iterasi dan meningkatkan kualitas klaster. Penelitian ini menggunakan dua dataset berbeda dari *UCI Machine Learning Repository*, yaitu *Whoscale Customer Dataset* dan QSAR Dataset. Kedua dataset ini dipilih untuk menguji kinerja algoritma *K-Medoids* dengan inisiasi *Z-Score* pada berbagai jenis data. Performa klastering dievaluasi menggunakan *Davies-Bouldin Index (DBI).*

*Davies-Bouldin Index (DBI)* merupakan salah satu metrik evaluasi yang digunakan untuk mengukur kualitas klaster yang dihasilkan oleh algoritma klastering (Nurdin, et al., 2022). DBI mengevaluasi rasio antara jarak dalam klaster dan jarak antar klaster, di mana semakin rendah nilai DBI, semakin baik kualitas klaster tersebut. Penggunaan DBI dalam evaluasi hasil klastering sangat penting karena mampu memberikan indikasi seberapa efektif data dikelompokkan, terutama dalam membandingkan beberapa algoritma klastering atau metode inisiasi yang berbeda (Asriningthias, et al., 2022). Penelitian ini, DBI digunakan untuk mengevaluasi kinerja algoritma *K-Medoids* dengan inisiasi medoids menggunakan Z-Score, guna memastikan peningkatan akurasi klaster yang dihasilkan dan membandingkan hasilnya dengan pendekatan lain.

Keunggulan dan kebaruan dari penelitian ini terletak pada penggabungan metode *Z-Score* untuk inisiasi medoids, yang menawarkan pendekatan inovatif dalam optimasi algoritma *K-Medoids*. Selain itu, penelitian ini juga menguji dua dataset yang berbeda untuk membuktikan generalisasi metode yang diusulkan, serta menggunakan *Davies-Bouldin Index (DBI)* sebagai metrik evaluasi untuk memberikan analisis yang lebih mendalam mengenai kualitas kluster yang dihasilkan.

Penelitian ini menjadi sangat penting karena adanya kebutuhan untuk meningkatkan efisiensi dan akurasi algoritma *K-Medoids* dalam menangani data yang besar dan kompleks. Dengan pesatnya perkembangan *machine learning* di berbagai sektor, seperti analisis data dan pengambilan keputusan bisnis, diperlukan algoritma yang mampu mengelola data secara optimal tanpa membebani komputasi. Penerapan *K-Medoids* yang optimal diharapkan memberikan keuntungan signifikan dalam kualitas hasil *clustering* dan pengambilan keputusan berbasis data. Inisiasi medoids yang tepat sangat krusial untuk mengurangi iterasi dan meningkatkan stabilitas hasil. Metode *Z-Score* sebagai pendekatan baru diharapkan dapat mengidentifikasi pusat cluster lebih akurat, mengurangi risiko hasil *clustering* yang tidak valid. Penelitian ini juga bertujuan memperkaya literatur tentang *K-Medoids* dengan memberikan perspektif baru mengenai optimasi teknik statistik, sehingga diharapkan dapat memberikan kontribusi berarti bagi

pengembangan teknik clustering di berbagai bidang, termasuk kesehatan, pemasaran, dan analisis sosial.

## 1.2. Rumusan Masalah

Salah satu permasalahan kritis dalam algoritma klastering *K-Medoids* adalah inisiasi medoids awal yang tidak optimal, terutama ketika diterapkan pada dataset berukuran besar. Inisiasi medoids yang kurang tepat dapat mengakibatkan peningkatan jumlah iterasi yang diperlukan untuk mencapai konvergensi dan menghasilkan klaster yang tidak representatif. Hal ini secara langsung mempengaruhi prediktabilitas dan stabilitas algoritma dalam menghasilkan klaster yang berkualitas. Oleh karena itu, dibutuhkan metode yang efektif untuk mengoptimalkan inisiasi medoids guna meningkatkan kinerja algoritma *K-Medoids*, baik dari segi efisiensi jumlah iterasi maupun kualitas klaster yang dihasilkan.

## 1.3. Tujuan Penelitian

Penelitian ini diharapkan bertujuan untuk:

1. Meningkatkan dengan metode *Z-Score* sebagai inisiasi medoids, sehingga dapat mengurangi jumlah iterasi clustering.
2. Meningkatkan kualitas klaster pada dataset besar dengan pemilihan medoids awal yang lebih tepat menggunakan normalisasi *Z-Score*.
3. Mengevaluasi performa algoritma menggunakan *Davies-Bouldin Index (DBI)* untuk memastikan *cluster* yang lebih valid dan optimal.

## 1.4 Manfaat Penelitian

Penelitian ini diharapkan memberikan manfaat yaitu:

1. Meningkatkan kualitas data dalam klastering sesudah dioptimasi dengan *z-score.*
2. Mengetahui kinerja *z-score* pada Algoritma *K-medoids* melalui perhitungan nilai performansi *DBI.*

3. Mengetahui nilai performansi *DBI* dan jumlah iterasi pada k-medoids konvensional.
4. Mengetahui nilai performansi *DBI* dan jumlah iterasi pada k-medoids + *z-score.*

## 1.5 Ruang Lingkup Batasan Masalah

Rumusan masalah dibatasi dengan beberapa hal berikut:

1. *Dataset* yang digunakan dalam penelitian ini yaitu:

a. *QSAR fish toxicity Dataset,* yang berjumlah 908 data dengan 7 atribut numeric yang diperoleh dari *UCI Machine Learning Repository.*

b. *Dataset Whoscale Qostumer,* yang berjumlah 440 data dengan 8 atribut numerik yang diperoleh dari *UCI Machine Learning Repository.*

2. Metode *Machine learning* yang digunakan adalah metode klastering.
3. Metode Inisiasi medoids yang digunakan adalah *Z-Score.*
4. Algoritma klastering yang digunakan adalah *K-Medoids.*
5. Metode evaluasi klastering yang digunakan adalah *Davies bouldin index.*