BAB II
TINJAUAN PUSTAKA

2.1 Penelitian Terdahulu yang Relevan

Sebagai bahan perbandingan dengan penelitian yang akan dilakukan, bagian ini menyajikan beberapa studi terdahulu yang membahas penerapan algoritma *clustering*, khususnya K-Means, dalam analisis data perbankan maupun sektor lain. Selain itu, ditampilkan juga penelitian yang menggunakan Silhouette Coefficient sebagai metode validasi dalam menentukan jumlah klaster yang optimal. Referensi yang digunakan berasal dari jurnal-jurnal yang terindeks MDPI, ScienceDirect, dan Google Scholar**,** yang kemudian dirangkum dalam tabel berikut.

2.1.1 Metrik Penelitian

Tabel 2. 1 Metrik Penelitian

1.	Nama Peneliti	Fatimah Defina Setiti Alhamdani , Ananda Ayu Dianti, Yufis Azhar
	Judul Jurnal	Segmentasi Pelanggan Berdasarkan Perilaku Penggunaan Kartu Kredit Menggunakan Metode K-Means Clustering
	Jenis Data	Historis transaksi kartu kredit
	Sumber Data	Dataset publik Kaggle (9000 data pengguna aktif kartu kredit)
	Metode	K-Means, Agglomerative Clustering, GMM (Gaussian Mixture Model), dan DBSCAN (Density-Based Spatial Clustering of Applications with Noise).
	Hasil	Hasil penelitian menunjukkan bahwa metode K-Means menghasilkan pengelompokan yang paling baik dibandingkan metode lainnya. Dengan menggunakan 9000 data pengguna aktif kartu kredit yang memiliki 18 variabel karakteristik, diperoleh tiga cluster utama: (1) nasabah dengan perilaku penggunaan kartu kredit yang rendah, (2) nasabah dengan penggunaan moderat, dan (3) nasabah dengan tingkat penggunaan yang tinggi dan sering melakukan

		transaksi. Nilai **Silhouette Coefficient** pada metode K-Means tercatat sebesar 0,207, yang meskipun relatif kecil, tetap lebih baik daripada Agglomerative, GMM, maupun DBSCAN. Hasil ini memperlihatkan bahwa K-Means mampu memberikan gambaran segmentasi pelanggan yang cukup jelas sehingga dapat dijadikan dasar dalam penyusunan strategi pemasaran kartu kredit secara lebih terarah.
	Model Evaluasi	Silhouette Coefficient (0,207), Elbow Method
	Tahun	2021
2.	Nama Peneliti	Meshal Shutaywi & Nezamoddin N. Kachouie
	Judul Jurnal	Silhouette Analysis for Performance Evaluation in Machine Learning with Applications to Clustering
	Jenis Data	Data benchmark (Bensaid, Dunn, Iris, Seed)
	Sumber Data	Benchmark dataset publik
	Metode	Kernel K-Means, Weighted Clustering
	Hasil	Mengusulkan metode **Weighted Clustering** berbasis **Silhouette Index** untuk memilih kernel terbaik; hasil menunjukkan metode ini mampu meningkatkan atau setidaknya menyamai performa kernel terbaik secara individual.
	Model Evaluasi	Silhouette Index, True Rate
	Tahun	2021
3.	Nama Peneliti	Aggelos Semoglou, Aristidis Likas, John Pavlopoulos
	Judul Jurnal	Silhouette-Guided Instance-Weighted k-means
	Jenis Data	Synthetic data & real-world datasets (Iris, Mice Protein, Glass, Wine, Digits, 20 Newsgroups)
	Sumber Data	Public datasets (UCI Repository, scikit-learn)
	Metode	K-Sil (Silhouette-Guided Instance-Weighted K-Means) dibandingkan dengan K-Means, DensityKMeans, dan LOFKMeans
	Hasil	K-Sil menghasilkan klaster yang lebih kohesif, robust terhadap noise/outlier, dan memberikan peningkatan

		signifikan pada kualitas klaster dibanding metode lain, khususnya dalam data dengan ketidakseimbangan dan noise.
	Model Evaluasi	Silhouette Score (micro & macro), NMI (Normalized Mutual Information), Wilcoxon signed-rank test
	Tahun	2025
4.	Nama Peneliti	Akash Punhani, Neetu Faujdar, Krishna Kumar Mishra, Manoharan Subramanian
	Judul Jurnal	Binning-Based Silhouette Approach to Find the Optimal Cluster Using K-Means
	Jenis Data	Data benchmark clustering (A-sets, S-sets, Birch dataset, UCI ML Repository)
	Sumber Data	Fundamental Clustering Problem Suite, UCI Repository
	Metode	K-Means dengan pendekatan binning + Silhouette Coefficient berbasis Central Limit Theorem (divide & conquer)
	Hasil	Metode yang diajukan mempercepat pencarian jumlah cluster optimal hingga **2,5–3 kali lebih cepat** dibanding metode iteratif tradisional, tanpa mengurangi kualitas hasil clustering.
	Model Evaluasi	Silhouette Score (internal validation), analisis waktu eksekusi
	Tahun	2022
5.	Nama Peneliti	Kanika Goyal, Deepak Garg
	Judul Jurnal	Segmentation of Credit Card Customers Based on Their Credit Card Usage Behavior using The K-Means Algorithm
	Jenis Data	Historis data transaksi kartu kredit
	Sumber Data	Dataset publik Kaggle (8950 pengguna kartu kredit)
	Metode	K-Means Clustering

	Hasil	Terbentuk 7 cluster perilaku pengguna kartu kredit, yang dapat digunakan untuk strategi pemasaran dan retensi pelanggan
	Model Evaluasi	Elbow Method, Silhouette Coefficient
	Tahun	2021
6.	Nama Peneliti	Wina Dwi Astuti
	Judul Jurnal	Algoritma K-Means untuk Meningkatkan Silhouette Score pada Pengelompokan Data Stok Bahan Manufaktur di PT. XYZ Kabupaten Majalengka
	Jenis Data	Data Stok Bahan Manufaktur
	Sumber Data	PT. XYZ Kabupaten Majalengka
	Metode	K-Means Clustering
	Hasil	Menghasilkan 3 cluster stok bahan: tinggi, sedang, rendah. Peningkatan nilai Silhouette Score dicapai dengan optimasi jumlah cluster.
	Model Evaluasi	Silhouette Score
	Tahun	2023
7.	Nama Peneliti	Siti Nur Illah, Nana Suarna, Irfan Ali, Dodi Solihudin
	Judul Jurnal	K-Means Clustering Method to Make Credit Payment Grouping Efficient
	Jenis Data	Data kredit (ID, Kecamatan, Kelurahan, Tenor, Last Payment)
	Sumber Data	Histori
	Metode	K-Means dengan proses KDD. Optimasi cluster dengan Davies-Bouldin Index (DBI).
	Hasil	Penelitian ini menghasilkan dua cluster optimal (K=2) dengan nilai Davies-Bouldin Index (DBI) 0.199, yang menandakan kualitas cluster yang cukup baik (semakin rendah nilai DBI, semakin baik pemisahan cluster). Cluster pertama merepresentasikan nasabah dengan tingkat pembayaran tinggi serta tenor pinjaman lebih panjang, menunjukkan profil debitur yang relatif stabil dan memiliki daya bayar yang kuat. Sebaliknya, cluster

		kedua berisi nasabah dengan pembayaran rendah serta tenor pendek, yang berpotensi menunjukkan adanya keterbatasan kemampuan bayar atau risiko gagal bayar lebih tinggi. Hasil ini dapat membantu lembaga keuangan dalam melakukan profiling risiko kredit, sehingga keputusan pemberian pinjaman bisa lebih selektif dan tepat sasaran.
	Model Evaluasi	Davies-Bouldin Index (DBI)
	Tahun	2025
8.	Nama Peneliti	Suryadi Muzahidi Aziz, Nur Azizah Komara Rifai
	Judul Jurnal	Pengelompokkan Ekspor Kopi Menurut Negara Tujuan Menggunakan Metode K-Means Clustering dengan Silhouette Coefficient
	Jenis Data	Data ekspor kopi (Berat Netto, Nilai FOB) per negara tujuan
	Sumber Data	Badan Pusat Statistik (BPS) Indonesia
	Metode	K-Means. Optimasi jumlah cluster hanya dengan Silhouette Coefficient.
	Hasil	Penelitian menemukan bahwa jumlah cluster optimal adalah K=2 dengan nilai Silhouette Coefficient 0.73, yang tergolong sangat baik (strong structure). Hal ini berarti pemisahan antar cluster sangat jelas. Cluster pertama berisi negara-negara dengan volume ekspor kopi yang tinggi seperti Jepang, Amerika Serikat, dan Jerman, yang menjadi pasar utama dan potensial. Sedangkan cluster kedua mencakup negara-negara lain dengan volume ekspor yang relatif rendah. Temuan ini menegaskan bahwa Indonesia memiliki pasar utama yang terkonsentrasi pada negara-negara maju dengan permintaan tinggi, sementara negara lain hanya berperan sebagai pasar pelengkap. Informasi ini sangat berguna bagi pemerintah maupun eksportir kopi untuk memperkuat hubungan dagang dengan cluster pasar

		utama sekaligus mengembangkan strategi penetrasi di pasar dengan permintaan rendah.
	Model Evaluasi	Silhouette Coefficient
	Tahun	2022
9.	Nama Peneliti	Yulia Indriani, Tria Setyani, Heni Sulistiani
	Judul Jurnal	Perbandingan Kinerja Algoritma K-Means Dan Dbscan Dalam Segmentasi Nasabah Berdasarkan Data Pemasaran Bank
	Jenis Data	Data profil dan interaksi nasabah bank (age, job, balance, dll.)
	Sumber Data	Dataset Bank Marketing dari Kaggle (https://www.kaggle.com/datasets/roundkbank/bank-marketing)
	Metode	Perbandingan K-Means vs DBSCAN. Evaluasi dengan Silhouette Score dan DBI.
	Hasil	Penelitian ini menunjukkan bahwa K-Means lebih unggul dibanding DBSCAN dalam segmentasi nasabah bank. Hasil evaluasi dengan Silhouette Score 0.231 dan DBI 1.295 pada K-Means lebih baik daripada DBSCAN yang hanya mencapai Silhouette Score 0.218 dan DBI 1.455. Walaupun nilai Silhouette masih rendah, perbedaan ini menunjukkan bahwa K-Means mampu membentuk cluster yang lebih jelas dengan variabilitas internal yang lebih rendah. Cluster yang dihasilkan dari K-Means dapat merepresentasikan kelompok nasabah dengan karakteristik berbeda seperti usia, pekerjaan, dan saldo tabungan. Hal ini relevan bagi bank untuk menyusun strategi pemasaran dan produk keuangan berbasis profil nasabah. Sementara DBSCAN kurang cocok karena data pemasaran bank cenderung memiliki distribusi yang seragam, sehingga pendekatan berbasis density kurang efektif.
	Model Evaluasi	Silhouette Score dan Davies-Bouldin Index (DBI)

	Tahun	2025
10.	Nama Peneliti	Putri Vania, Betha Nurina Sari
	Judul Jurnal	Perbandingan Metode Elbow dan Silhouette untuk Penentuan Jumlah Klaster yang Optimal pada Clustering Produksi Padi menggunakan Algoritma K-Means
	Jenis Data	Data luas dan produksi padi di Jawa Barat pada tahun 2020,2021, dan 2022.
	Sumber Data	Badan Pusat Statistik (BPS) Jawa Barat
	Metode	Elbow vs Silhouette untuk menentukan K optimal dalam K-Means. Evaluasi akhir dengan DBI.
	Hasil	Penelitian membandingkan hasil penentuan jumlah cluster menggunakan Elbow Method dan Silhouette Coefficient dalam analisis produksi padi. Elbow menyarankan jumlah cluster K=3 dengan DBI=0.39, sedangkan Silhouette menyarankan K=5 dengan DBI=0.27. Karena DBI yang lebih rendah menunjukkan kualitas cluster yang lebih baik, maka hasil dari metode Silhouette dianggap lebih akurat. Pada K=5, data produksi padi dan luas lahan terbagi lebih detail sehingga mampu menangkap variasi antar daerah di Jawa Barat. Hal ini memungkinkan pemerintah daerah untuk menyusun kebijakan pertanian yang lebih spesifik, misalnya memberikan bantuan pupuk di daerah dengan hasil rendah, serta memperkuat distribusi di daerah dengan produksi tinggi. Dengan demikian, penggunaan Silhouette lebih unggul dibanding Elbow dalam menghasilkan clustering yang benar-benar representatif.
	Model Evaluasi	Davies-Bouldin Index (DBI)
	Tahun	2023

11.	Nama Peneliti	Marcel-Ioan Bolog, Stefan Rusu, Marius Leordeanu, Claudia Diana Sabău-Popa, Diana Claudia Perticas, Mihai-Ioan Crisan
	Judul Jurnal	K-Means Clustering for Portfolio Optimization: Symmetry in Risk–Return Tradeoff, Liquidity, Profitability, and Solvency
	Jenis Data	Data finansial kuartalan perusahaan (profitabilitas, likuiditas, solvabilitas).
	Sumber Data	Data fundamental: Calcbench API (dari laporan SEC).
	Metode	K-Means, Elbow Method dan Silhouette Analysis
	Hasil	Portofolio pada Cluster 1 yang berisi perusahaan dengan profil lebih berisiko mampu memberikan return lebih tinggi, namun diikuti dengan tingkat volatilitas yang juga lebih besar. Sebaliknya, Cluster 0 menunjukkan karakteristik yang lebih stabil dengan keseimbangan risk-return yang moderat. Hasil analisis juga menunjukkan bahwa portofolio yang dibentuk berdasarkan indikator ROA, OCFM, dan GPM menghasilkan Sharpe Ratio tertinggi, sehingga memberikan kinerja yang lebih baik dibandingkan benchmark pasar. Selain itu, penelitian ini menegaskan bahwa Silhouette Analysis lebih andal dibandingkan Elbow Method dalam menentukan jumlah cluster yang optimal, sehingga mendukung akurasi pengelompokan.
	Model Evaluasi	Sharpe Ratio sebagai ukuran utama kinerja risk-adjusted, dengan tambahan Annual Return (AR), Annual Volatility (AV), serta perbandingan terhadap benchmark SPY.
	Tahun	2025

Penelitian mengenai risiko kredit dan segmentasi nasabah dengan algoritma *clustering* sudah cukup banyak dilakukan, namun masih terdapat celah penelitian (*research gap*) yang perlu diperhatikan. Beberapa penelitian terdahulu, seperti yang

dilakukan oleh (Bakoben et al., 2017) dalam *Identification of Credit Risk Based on Cluster Analysis of Account Behaviours* menunjukkan bahwa analisis cluster mampu mendeteksi pola perilaku risiko kredit nasabah. Namun, penelitian ini lebih menekankan pada perilaku akun secara umum dan belum secara spesifik mengintegrasikan indikator finansial utama seperti total pokok pinjaman, rate, income material, dan nilai CKPN yang relevan untuk konteks perbankan lokal.

Selanjutnya, studi dari IJACSA (2021) berjudul *Machine Learning Mini Batch K-Means and Business Intelligence Utilization for Credit Card Customer Segmentation* juga berhasil memanfaatkan algoritma *K-Means* dalam mengelompokkan pelanggan kartu kredit. Akan tetapi, fokusnya lebih kepada perilaku transaksi kartu kredit, sehingga kurang relevan apabila langsung diterapkan pada data nasabah kredit perbankan daerah, khususnya Bank MalukuMalut.

Di sisi lain, penelitian open access oleh MDPI (2022), *K-Means Clustering Approach for Intelligent Customer Segmentation* menunjukkan bahwa metode K-Means dapat dimanfaatkan secara efektif dalam segmentasi berbasis data numerik. Meski demikian, penelitian tersebut masih terbatas pada sektor e-commerce dan belum diterapkan pada domain risiko kredit perbankan. Dengan demikian, terdapat kesenjangan penelitian bahwa belum ada studi yang secara khusus menerapkan K-Means Clustering pada data nasabah kredit Bank MalukuMalut dengan indikator finansial dan demografis lokal, serta melakukan validasi cluster menggunakan Silhouette Coefficient.

Sejalan dengan itu, penelitian ini bertujuan untuk mengidentifikasi pola dan memprofil risiko kredit nasabah Bank MalukuMalut dengan menggunakan algoritma K-Means Clustering**.** Penelitian ini berfokus pada pengelompokan nasabah berdasarkan karakteristik finansial dan demografis yang relevan, sehingga bank dapat menilai risiko kredit secara lebih akurat. Proses evaluasi cluster dilakukan menggunakan Silhouette Coefficient, yang berfungsi untuk menentukan jumlah cluster (K) optimal serta menilai seberapa baik pemisahan antar cluster. Dengan pendekatan ini, setiap cluster yang terbentuk diharapkan dapat mencerminkan tingkat risiko kredit yang berbeda, mulai dari risiko rendah, sedang, hingga tinggi.

a. Metode yang digunakan:

Penelitian sebelumnya banyak menerapkan K-Means Clustering dan variasinya (seperti K-Sil, Kernel K-Means, atau K-Means yang dioptimasi dengan Davies-Bouldin Index), bahkan beberapa membandingkannya dengan metode lain seperti Agglomerative Clustering, Gaussian Mixture Model (GMM), maupun DBSCAN. Evaluasi yang digunakan antara lain Silhouette Score, Elbow Method, Davies-Bouldin Index (DBI), dan ukuran performa lain seperti Sharpe Ratio pada kasus portofolio.

Berbeda dengan penelitian ini, penulis secara khusus menggunakan algoritma K-Means Clustering untuk profiling risiko kredit nasabah Bank MalukuMalut, dengan fokus evaluasi melalui Silhouette Coefficient untuk menentukan jumlah cluster optimal. Hal ini menjadikan penelitian lebih terarah pada konteks perbankan dan pengelolaan risiko kredit, bukan hanya pada segmentasi umum atau data benchmark.

b. Dataset yang digunakan:

Penelitian sebelumnya umumnya mengandalkan dataset publik seperti data transaksi kartu kredit dari Kaggle, data benchmark (Iris, Wine, Digits, dll.), data pertanian dari BPS, maupun data finansial global. Fokus analisis lebih ke segmentasi pelanggan, perilaku transaksi, atau pembentukan portofolio. Sementara itu, penelitian ini menggunakan data historis sekunder nasabah kredit Bank MalukuMalut, yang mencakup variabel-variabel penting terkait profil nasabah dan karakteristik kreditnya. Fokus penelitian diarahkan untuk mengidentifikasi karakteristik risiko kredit dengan pendekatan clustering, sehingga dapat membantu bank dalam membuat profil risiko nasabah serta mendukung pengambilan keputusan yang lebih tepat dalam pemberian maupun pengelolaan kredit.

2.2 Landasan Teori

Untuk mendasari pelaksanaan penelitian ini, akan dibahas berbagai teori yang relevan. Berikut ini penulis akan menjelaskan teori-teori yang menjadi dasar dalam penelitian ini.

2.2.1 Clustering

Clustering merupakan metode dalam *data mining* yang digunakan untuk mengelompokkan data ke dalam beberapa klaster berdasarkan kemiripan ciri atau atribut. Data yang memiliki karakteristik serupa akan ditempatkan dalam klaster yang sama, sedangkan data dengan karakteristik berbeda dikelompokkan ke dalam klaster lain. Tujuan utama clustering adalah membentuk klaster yang homogen di dalamnya serta memperjelas perbedaan antar klaster. Penerapan clustering pada data gangguan dapat membantu mengidentifikasi pola lokasi yang sering bermasalah sehingga upaya pencegahan dan perbaikan dapat dilakukan secara lebih efektif (Amri et al., 2024).

Clustering juga merupakan salah satu metode pengelompokan data non-hierarkis yang bertujuan membagi data ke dalam satu atau lebih klaster melalui optimasi fungsi objektif, dengan meminimalkan variasi di dalam klaster dan memaksimalkan variasi antar klaster. Secara umum, algoritma clustering terdiri dari beberapa kategori, salah satunya adalah metode partisi, di mana jumlah klaster (K) ditentukan di awal dan setiap data hanya dapat menjadi anggota satu klaster tanpa adanya tumpang tindih, sebagaimana diterapkan pada algoritma K-Means (Arianti et al., 2018).

2.2.2 Risiko Kredit

Risiko kredit adalah potensi kerugian yang dialami bank ketika debitur tidak mampu menunaikan kewajiban pembayaran pokok atau bunga sesuai perjanjian. Risiko ini dapat terlihat dalam berbagai situasi, misalnya penundaan pelunasan, tidak adanya pembayaran sama sekali, ataupun turunnya nilai jaminan yang diberikan. Karena sifatnya yang langsung memengaruhi stabilitas keuangan, risiko kredit menjadi salah satu tantangan utama perbankan yang dapat mengganggu kesehatan sistem keuangan apabila tidak dikelola secara efektif (Silitonga & Manda, 2022).

Secara umum, risiko kredit dipahami sebagai risiko kerugian akibat pihak peminjam tidak dapat atau tidak bersedia melunasi kewajiban pinjaman pada waktu yang telah ditetapkan. Risiko ini tidak dapat dihindari karena salah satu fungsi pokok bank adalah menyalurkan dana dalam bentuk kredit kepada masyarakat. Untuk mengukur tingkat risiko kredit, perbankan biasanya menggunakan indikator Non-Performing Loan (NPL). Rasio NPL mencerminkan kualitas manajemen dalam menangani kredit bermasalah. Semakin tinggi nilai NPL, semakin buruk kualitas kredit yang disalurkan,

sehingga jumlah pinjaman bermasalah meningkat. Kondisi tersebut dapat menekan pendapatan bunga serta laba bank, mengingat bank tetap berkewajiban membayar bunga simpanan kepada nasabah meskipun kredit bermasalah tidak tertagih (Silitonga & Manda, 2022).

2.2.3 Non-Performing Loan (NPL)

Pada dasarnya, kredit merupakan bentuk perjanjian pinjaman antara dua pihak, di mana debitur berkewajiban mengembalikan dana pinjaman sesuai jangka waktu yang telah ditetapkan, disertai bunga atau biaya lain yang disepakati. Dalam penerapannya, salah satu risiko utama yang sering muncul adalah ketidakmampuan debitur dalam memenuhi kewajiban pembayaran. Situasi ini dikenal sebagai kredit bermasalah atau Non-Performing Loan (NPL) (Sari Maya et al., 2021).

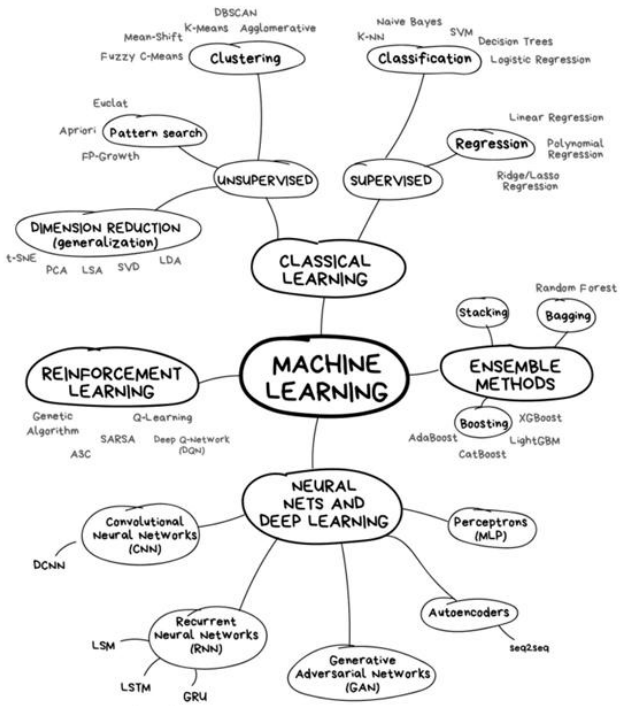
Kredit bermasalah menggambarkan kondisi ketika debitur kesulitan melunasi kewajibannya, baik sebagian maupun secara penuh, sebagaimana yang tercantum dalam perjanjian pinjaman. Peningkatan rasio NPL menunjukkan menurunnya tingkat kesehatan perbankan. Hal ini tidak hanya menurunkan kepercayaan masyarakat terhadap bank, tetapi juga berpotensi mengurangi tingkat profitabilitas lembaga keuangan tersebut(Sari Maya et al., 2021).

2.2.4 Machine Learning

Machine Learning saat ini telah menjadi salah satu elemen penting dalam berbagai model bisnis modern, contohnya pada sistem pencarian Google. Teknologi ini tidak hanya digunakan di perusahaan berbasis teknologi, tetapi juga mulai diadopsi secara luas oleh berbagai sektor industri untuk mendukung aktivitas operasional. Seiring dengan meningkatnya kebutuhan dan ketertarikan terhadap teknologi ini, penelitian serta praktik terus berkembang guna menemukan metode implementasi yang lebih optimal (Suprihadi Eddy, 2022)..

Sebagai bagian dari Kecerdasan Buatan (Artificial Intelligence/AI), Pembelajaran Mesin memiliki keunikan karena tidak mengandalkan instruksi pemrograman secara langsung. Konsep dasarnya pertama kali diperkenalkan oleh Alan Turing pada tahun 1950-an, kemudian mendapatkan perhatian yang lebih luas pada dekade 1990-an. Popularitasnya semakin melonjak setelah keberhasilan DeepMind anak perusahaan Google mengalahkan juara dunia Go pada tahun 2016. Sejak peristiwa tersebut,

penerapan dan pengembangan Pembelajaran Mesin semakin meluas di berbagai bidang kehidupan (Suprihadi Eddy, 2022).



Gambar 2. 1 Map Machine Learning (Suprihadi Eddy, 2022)

Salah satu kekuatan utama pembelajaran mesin adalah kemampuannya dalam mengolah serta menganalisis data berukuran besar dengan kecepatan dan efisiensi tinggi. Informasi yang diperoleh dari hasil analisis tersebut dapat dijadikan dasar untuk mendukung pengambilan keputusan yang lebih tepat. Arthur Samuel pada tahun 1959 mendeskripsikan pembelajaran mesin sebagai cabang dari ilmu komputer yang memungkinkan sistem komputer untuk "belajar" tanpa perlu instruksi pemrograman secara langsung, sehingga menjadikannya teknologi yang sangat penting di era big data saat ini (Suprihadi Eddy, 2022).

2.2.5 Davies-Bouldin Index (DBI)

Davies-Bouldin Index (DBI) merupakan salah satu metode evaluasi internal yang digunakan untuk menilai validitas hasil clustering. Indeks ini bekerja dengan meminimalkan jarak antar data dalam satu klaster (*intra-cluster distance*) dan memaksimalkan jarak antar klaster (*inter-cluster distance*). Semakin kecil jarak di dalam klaster, maka tingkat kemiripan antar objek dalam klaster tersebut semakin tinggi,

sedangkan semakin besar jarak antar klaster menunjukkan pemisahan kelompok yang semakin jelas. Dengan demikian, DBI memberikan ukuran kuantitatif terhadap kualitas hasil clustering dalam membentuk klaster yang kompak (*cohesive*) dan terpisah (*separated*) (Mughnyanti et al., 2020)

Dalam penelitian ini, Davies-Bouldin Index (DBI) digunakan untuk menentukan nilai K yang optimal pada proses clustering. Penilaian dilakukan dengan membandingkan nilai DBI pada beberapa skenario jumlah klaster, di mana nilai DBI yang paling rendah menunjukkan hasil pengelompokan yang paling optimal. Nilai DBI yang kecil mencerminkan klaster yang memiliki tingkat kemiripan internal yang tinggi serta pemisahan antar klaster yang jelas (Asri et al., 2020).

Rumus DBI dapat dituliskan sebagai berikut:

$$DBI = \frac{1}{k}\sum_{i=1}^{k} max\, j \neq i \frac{Si + Sj}{M_{ij}} \qquad \textbf{(2. 1)}$$

dengan:

a) k = jumlah cluster,
b) Si = rata-rata jarak setiap titik dalam cluster i terhadap centroid-nya (ukuran kompakness intra-cluster),
c) M_{ij} = jarak antara centroid cluster i dan cluster j (ukuran separation antar cluster).

Interpretasi nilai DBI:

a) Semakin kecil nilai DBI → semakin baik kualitas clustering, karena cluster yang terbentuk semakin rapat di dalamnya dan semakin jelas perbedaannya dengan cluster lain.
b) Nilai DBI besar → clustering kurang baik, karena cluster saling tumpang tindih dan batas antar cluster kurang tegas.

2.2.6 K-Means

K-Means adalah metode klasterisasi non-hierarkis yang berfungsi untuk membagi data ke dalam dua atau lebih kelompok, dengan tujuan mengelompokan data yang memiliki kesamaan ke dalam klaster yang sama (Setyani et al., 2025). Algoritma ini bekerja dengan membagi data ke dalam *K* kelompok berdasarkan kedekatan jarak, yang umumnya dihitung menggunakan *Euclidean distance*. K-Means merupakan algoritma

clustering yang digunakan untuk mengelompokkan data tanpa label ke dalam beberapa klaster berdasarkan tingkat kemiripan. Dalam penelitian ini, K-Means digunakan untuk mengelompokkan instruktur berdasarkan materi diklat, sedangkan metode TOPSIS digunakan untuk menentukan instruktur terbaik pada setiap klaster berdasarkan kedekatannya dengan solusi ideal (Dyah Budiana et al., 2019).

Secara umum, K-Means merupakan salah satu metode *data clustering* non-hierarkis yang berfungsi mempartisi data ke dalam satu atau lebih klaster dengan mengelompokkan data yang memiliki karakteristik serupa ke dalam klaster yang sama. Penerapan K-Means pada data keluhan pelanggan dapat membantu mengelompokkan keluhan yang sejenis sehingga memudahkan proses analisis dan identifikasi permasalahan yang sering terjadi. Melalui pendekatan *data mining*, proses ini bertujuan untuk menemukan pola serta informasi penting dari kumpulan data dalam jumlah besar *(Hartanti, 2015)*.

Tahapan utama dalam K-Means meliputi: inisialisasi centroid, perhitungan jarak setiap data terhadap centroid, pembaruan posisi centroid, serta pengulangan proses tersebut hingga tercapai kondisi optimal atau konvergen.

Secara umum prosedur metode K-Means adalah sebagai berikut:

1. Tahap awal dalam algoritma K-Means adalah menentukan jumlah klaster (k) yang akan dibentuk. Nilai k dapat ditentukan secara manual berdasarkan kebutuhan analisis, atau menggunakan metode evaluasi seperti :
    a) Metode Elbow → melihat titik tekukan pada grafik *inertia* (Within-Cluster Sum of Squares) untuk menentukan K Optimal.
    b) Metode Silhouette → mengevaluasi kualitas klaster dengan melihat nilai *Silhouette Score* yang menunjukkan tingkat keseragaman dalam satu klaster dan keterpisahan antar klaster.
2. Inisialisasi centroid dilakukan secara acak dengan memilih sejumlah objek sesuai jumlah *k* klaster. Centroid awal ini menjadi dasar perhitungan centroid klaster ke-i pada iterasi berikutnya, dengan rumus:

$$c = \frac{\sum_{i=1}^{n} x_i}{n} \; ; i = 1,2,3, \dots n \qquad \textbf{(2. 2)}$$

Dimana :

a) c = centroid pada cluster
b) x_i = obyek ke - i
c) n = banyaknya obyek atau jumlah obyek yang menjadi anggota cluster

Rumus jarak (Euclidean Distance)

1. Hitung jarak setiap objek ke masing-masing centroiddari masing-masing cluster. Kemudian hitung jarak antara objek dengan centroid, dalam penelitian ini menggunakan Euclidian Distance:

$$d(x,y) = ||x-y|| = \sqrt{\sum_{i=1}^{n}(xi-yi)^2} \; ; i = 1,2,3,\dots n \qquad \textbf{(2. 3)}$$

Dimana:

a) xi : objek x ke-i
b) y: daya y ke-i
c) n: banyaknya objek

2. Pengelompokan objek ke klaster

Setiap objek ditempatkan pada klaster berdasarkan jarak terdekatnya dengan centroid yang sudah ditentukan.

3. Iterasi dan pembaruan centroid

Setelah objek masuk ke dalam klaster, dilakukan perhitungan ulang untuk menentukan posisi centroid baru berdasarkan rata-rata dari seluruh anggota klaster.

4. Proses berulang hingga konvergen

Langkah pengelompokan dan pembaruan centroid ini diulangi terus-menerus sampai posisi centroid tidak lagi berubah secara signifikan, sehingga klaster yang terbentuk dianggap stabil.

Proses pengelompokan titik dilakukan dengan membandingkan matriks kumpulan tugas pada iterasi sebelumnya dengan matriks kumpulan tugas pada iterasi saat ini. Jika hasil perbandingan menunjukkan kesamaan, maka algoritma K-Means telah mencapai konvergensi. Namun, jika terdapat perbedaan, proses iterasi perlu dilanjutkan hingga tercapai kondisi konvergen. Analisis hasil dilakukan dengan

mendeskripsikan karakteristik klaster yang terbentuk, suatu proses yang dikenal sebagai profilingcluster. Hasil yang diharapkan dari analisis ini adalah memperoleh informasi yang relevan mengenai segmentasi nasabahberdasarkan atribut yang digunakan dalam penelitian(Irawan et al., 2025) .

2.2.7 Silhouette Score

Silhouette merupakan metode evaluasi klaster yang menggabungkan dua konsep, yaitu cohesion dan separation. Cohesion diukur melalui kedekatan antar objek dalam satu klaster, sedangkan separation diukur berdasarkan jarak rata-rata suatu objek terhadap klaster terdekatnya. Penentuan jumlah klaster yang optimal sesuai dengan dataset sangat penting, karena akan memengaruhi kualitas hasil klasterisasi dan berimplikasi pada pengambilan keputusan. Oleh karena itu, diperlukan metode yang mampu menentukan jumlah klaster (K) yang tepat dan optimal. Dalam penelitian ini, metode Silhouette digunakan sebagai acuan untuk menentukan jumlah klaster terbaik (Rizqi Mulyawan et al., n.d.).

Evaluasi hasil clustering pada penelitian ini menggunakan Silhouette Score. Metrik ini menilai seberapa baik objek berada dalam klusternya sendiri dibandingkan dengan kedekatannya terhadap kluster lain. Nilai Silhouette Score berada pada rentang -1 hingga 1, di mana nilai mendekati 1 menunjukkan kualitas kluster yang baik, nilai mendekati 0 menunjukkan objek berada pada batas kluster, sedangkan nilai negatif mengindikasikan salah pengelompokan. Semakin tinggi nilai Silhouette Score, semakin baik kualitas clustering yang dihasilkan (Setyani et al., 2025).

Rumus Silhouette Score:

$$s(i) = \frac{\mathrm{b}(i) - a(i)}{\max\{a(i), b(i)\}} \qquad \textbf{(2. 4)}$$

Dengan keterangan sebagai berikut:

a) $s(i)$: Merupakan nilai Silhouette untuk data ke-i, yang menunjukkan seberapa baik titik tersebut cocok dengan klusternya.

b) $a(i)$: Menunjukkan rata-rata jarak antara titik data ke-i dengan seluruh titik lain dalam kluster yang sama (mengukur kohesi).

c) $b(i)$: Menunjukkan rata-rata jarak antara titik data ke-i dengan seluruh titik dalam kluster terdekat lainnya (mengukur pemisahan).

Setelah semua titik memiliki nilai s(i), maka dihitung rata-ratanya untuk mendapatkan Silhouette Score total:

Silhouette Score Total = (1/n) * Σ s(i)

Dimana:

-n: jumlah total titik data

Interpretasi:

+1 : Titik sangat cocok dalam klusternya.

0 : Titik berada di antara dua kluster.

-1 : Titik kemungkinan salah dikelompokkan.

2.2.8 Metode Elbow

Clustering merupakan salah satu teknik analisis data yang digunakan untuk mengelompokkan data ke dalam beberapa kelompok berdasarkan kemiripan karakteristik. Salah satu metode clustering yang paling populer adalah K-Means, karena memiliki kemampuan mengolah data dalam jumlah besar dengan relatif cepat dan efisien. Namun, kelemahan utama dari metode ini adalah penentuan jumlah cluster K yang optimal. Jika K tidak ditentukan dengan tepat, hasil pengelompokan dapat menjadi kurang akurat (Maori, 2023).

Untuk mengatasi hal tersebut, digunakan Elbow Method yang berfungsi membantu menentukan jumlah cluster terbaik. Metode ini bekerja dengan cara menghitung nilai Within-Cluster Sum of Squares (WCSS) untuk berbagai nilai K, kemudian memvisualisasikannya dalam bentuk grafik. Titik siku ("elbow point") pada grafik tersebut merepresentasikan jumlah cluster optimal, karena setelah titik itu penurunan nilai WCSS tidak lagi signifikan (Maori, 2023).

Rumus dasar WCSS dalam metode Elbow dituliskan sebagai berikut:

$$WCSS = \sum_{k=1}^{K} \sum_{x_i \in C_k} \left|\left|x_i - \mu_k\right|\right|^2 \qquad \textbf{(2. 5)}$$

dengan:

a) k = jumlah cluster,

b) C_i = cluster ke-i,

c) x = data anggota cluster,

d) μ_i = centroid cluster ke-i,

e) $\left|\left|x - \mu_i\right|\right|^2$ = jarak Euclidean antara data x dan centroid μ_i.

Interpretasi:

a) Semakin kecil nilai WCSS → semakin kompak cluster.

b) Jumlah K optimal ditentukan pada titik "siku" di grafik, di mana penurunan WCSS mulai melambat.

2.2.9 Min-Max Normalization

Normalisasi data merupakan salah satu tahap penting dalam proses pra-pengolahan data (data preprocessing) yang bertujuan untuk menyamakan skala nilai dari setiap atribut agar dapat dibandingkan secara proporsional. Hal ini diperlukan karena pada dataset sering kali terdapat atribut dengan rentang nilai yang sangat berbeda. Misalnya, atribut "pendapatan" dapat memiliki rentang jutaan, sedangkan atribut "jumlah tanggungan" mungkin hanya berkisar 0 sampai 10. Jika tidak dilakukan normalisasi, atribut dengan nilai besar akan mendominasi proses perhitungan, sehingga hasil analisis atau pemodelan menjadi bias (Selle et al., 2022).

Salah satu teknik normalisasi yang paling sederhana dan sering digunakan adalah Min-Max Normalization. Metode ini bekerja dengan melakukan transformasi linier pada data asli sehingga nilai atribut ditransformasikan ke dalam skala tertentu, biasanya 0 hingga 1. Dengan demikian, data hasil transformasi akan memiliki keseimbangan skala antar atribut, tanpa mengubah hubungan proporsional dari data aslinya (Selle et al., 2022). Secara matematis, rumus Min-Max Normalization dapat dituliskan sebagai berikut:

$$d' = \frac{d - min}{\max - min} \quad \textbf{(2. 6)}$$

dengan:

a) d' = nilai hasil normalisasi.
b) d = nilai data asli.
c) min = nilai terkecil pada atribut tertentu.
d) max = nilai terbesar pada atribut tersebut.

Interpretasi dari persamaan tersebut adalah setiap nilai asli d dikurangi dengan nilai minimum, kemudian dibagi dengan rentang ($max - min$). Hasilnya adalah nilai baru d' yang berada pada skala tertentu, umumnya antara 0 dan 1.

2.2.10 Bahasa Pemrograman Python

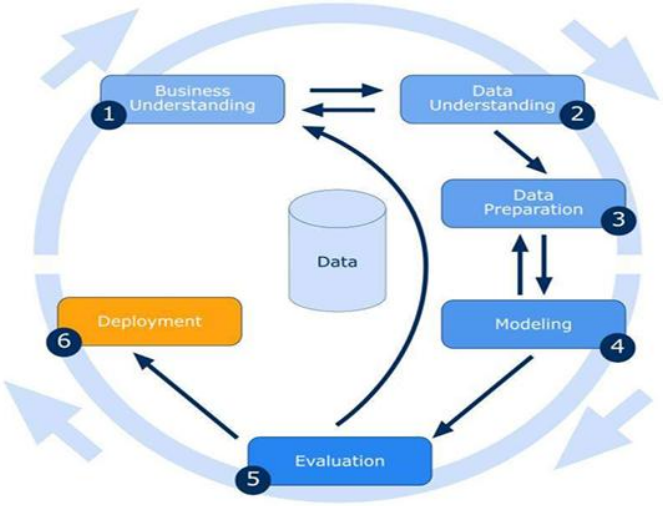
Python dikembangkan dengan mengambil inspirasi dari bahasa pemrograman ABC yang pernah populer pada masanya. Keunggulan utama Python dibandingkan bahasa pemrograman lainnya terletak pada sifatnya yang open source, sehingga perkembangannya melibatkan kontribusi jutaan pengembang, peneliti, dan pengguna dari berbagai bidang, tidak terbatas hanya pada sektor teknologi informasi. Python merupakan

bahasa berbasis interpreter, sehingga kode dapat dijalankan langsung tanpa melalui proses kompilasi, serta kompatibel dengan beragam sistem operasi seperti Windows, Linux, dan lain-lain. Selain itu, Python mendukung berbagai paradigma pemrograman, mulai dari prosedural seperti pada C, berorientasi objek seperti Java, hingga fungsional seperti Lisp. Fleksibilitas paradigma tersebut memberikan nilai tambah bagi programmer dalam membangun berbagai jenis aplikasi menggunakan Python (Rahman et al., n.d.).

2.2.11 Metodologi CRISP-DM

CRISP-DM (*Cross Industry Standard Process for Data Mining*) merupakan salah satu kerangka kerja standar yang paling banyak digunakan dalam praktik *data mining* karena menyediakan alur kerja yang sistematis, terstruktur, dan efisien. Kerangka ini membantu proses pengolahan data menjadi lebih konsisten, mudah direplikasi, serta mendukung pengelolaan proyek analitik yang kompleks. Data mining merupakan serangkaian proses yang bertujuan untuk menemukan dan mengidentifikasi pola dalam data, di mana pola yang dihasilkan bersifat valid, baru, bermanfaat, serta mudah dipahami sehingga dapat digunakan sebagai dasar pengambilan keputusan(Asri et al., 2015).

Penerapan CRISP-DM tidak hanya relevan di bidang bisnis, tetapi juga banyak digunakan dalam penelitian akademik karena mampu mengarahkan proses analisis secara lebih fokus dan terencana. Oleh karena itu, penelitian ini mengadopsi metodologi CRISP-DM sebagai pendekatan pemecahan masalah yang bersifat umum dan fleksibel, yang terdiri dari enam tahapan utama, yaitu *business understanding*, *data understanding*, *data preparation*, *modeling*, *evaluation*, dan *deployment* (Hasanah et al., 2021). Data mining merupakan proses untuk menggali dan mengungkap pengetahuan yang tersimpan di dalam basis data melalui pencarian pola, hubungan, serta nilai tambah dari kumpulan data berukuran besar. Proses ini bertujuan menyederhanakan data menjadi informasi yang bermakna dan mudah dipahami, sehingga dapat dimanfaatkan secara optimal dalam pengambilan keputusan. Dalam pelaksanaannya, data mining memanfaatkan pendekatan statistik dan matematika untuk menghasilkan informasi yang bernilai dan bermanfaat (Fitriyadi, 2021).



Gambar 2. 2 Metodologi CRISP-DM

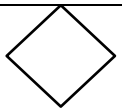
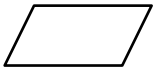
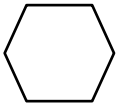
Gambar 2.2 menggambarkan enam tahapan inti dari CRISP-DM yang saling terkait. Penjelasan setiap tahapannya adalah sebagai berikut:

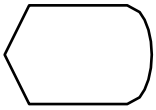
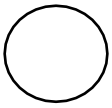
1. **Business Understanding**: Tahap awal yang menekankan pentingnya memahami secara mendalam tujuan serta kebutuhan bisnis, sekaligus mendefinisikan permasalahan yang akan diselesaikan melalui data mining.
2. **Data Understanding**: Melibatkan proses pengumpulan, eksplorasi, serta pemeriksaan kualitas dan karakteristik data untuk menemukan pola awal yang bermanfaat bagi analisis berikutnya.
3. **Data Preparation**: Proses pengolahan data agar siap digunakan dalam pemodelan, mencakup pembersihan, transformasi, dan pemilihan atribut sesuai kebutuhan analisis.
4. **Modeling**: Penerapan teknik atau algoritma pemodelan yang sesuai untuk melakukan prediksi maupun klasifikasi. Model yang dihasilkan kemudian diuji untuk menilai akurasi dan kinerjanya..
5. **Evaluation**: Mengevaluasi performa model dengan membandingkan hasilnya terhadap tujuan penelitian atau kebutuhan bisnis, sekaligus melakukan perbaikan bila diperlukan.
6. **Deployment**: Mengimplementasikan model ke dalam sistem operasional guna mendukung pengambilan keputusan. Tahap ini juga mencakup pemantauan serta pemeliharaan agar model tetap memberikan hasil optimal secara berkelanjutan.

2.2.12 Flowchart

Dengan perkembangan teknologi, penggunaan diagram alir semakin meluas karena mampu membantu menjelaskan algoritma dalam bentuk langkah-langkah sistematis yang dapat diterapkan ke dalam bahasa pemrograman. Diagram alir pada dasarnya merupakan representasi visual dari tahapan suatu algoritma yang disusun secara terstruktur, sehingga informasi yang disajikan menjadi lebih mudah dipahami dan komunikatif. Beberapa jenis diagram alir yang sering dipakai antara lain: flowchart sistem, flowchart dokumen, flowchart skematik, flowchart program, serta flowchart proses (Ni Nyoman Emang Smrti et al., n.d.).

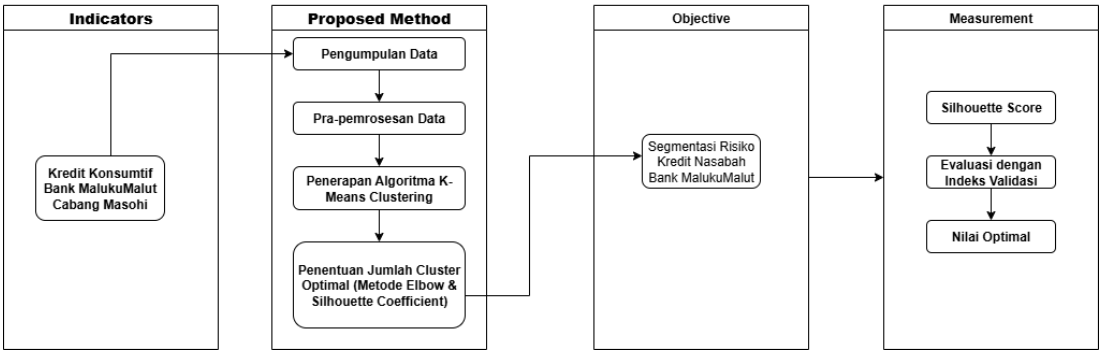
Tabel 2. 2 Flowchart (Ni Nyoman Emang Smrti et al., n.d.)

No.	Simbol	Nama	Keterangan
1.		*Terminator*	Digunakan untuk menandai awal (start) dan akhir (stop) dari suatu proses.
2.		*Flow Direction*	Digunakan untuk menghubungkan satu simbol dengan simbol lainnya.
3.		*Processing*	Simbol yang menunjukkan bahwa komputer sedang melakukan pengolahan data
4.		*Decision*	Simbol proses yang dipilih sesuai dengan kondisi yang ada saat ini.
5.		*Input – Output*	Simbol yang menunjukkan proses input dan output tanpa memandang jenis peralatan yang digunakan.
6.		*Preparation*	Simbol yang digunakan untuk menyiapkan data yang akan disimpan dalam penyimpanan untuk diproses.

7.		*Predefine Proses*	Simbol yang menggambarkan pelaksanaan komponen atau prosedur tertentu.
8.		*Display*	Simbol yang menggambarkan pelaksanaan komponen atau prosedur tertentu.
9.		*Punch Card*	Simbol yang menandakan bahwa output ditulis ke kartu atau input berasal dari kartu.
10.		*Connector*	Simbol yang digunakan untuk menyambungkan proses dalam satu lembar atau halaman yang sama.

2.2.13 Kerangka Pemikiran

Pada bagian ini dijelaskan alur penelitian yang dilaksanakan, dimulai dari proses pengumpulan data nasabah kredit di Bank Maluku Cabang Masohi. Selanjutnya dilakukan tahap pra-pengolahan data, penerapan algoritma K-Means Clustering, serta penentuan jumlah cluster yang paling sesuai menggunakan metode Elbow dan Silhouette Coefficient. Tahapan terakhir mencakup evaluasi serta interpretasi hasil pengelompokan untuk memetakan risiko kredit nasabah.



Gambar 2. 3 Kerangka Pemikiran

Gambar di atas menggambarkan proses penelitian yang akan dilakukan oleh penulis dalam studi ini, yang penjelasannya akan dibahas sebagai berikut:

1. ***Indicators***: Bagian ini berisi data nasabah kredit Bank MalukuMalut yang digunakan sebagai dataset utama untuk proses analisis dan pengelompokan risiko kredit.
2. ***Proposed Method***: Tahapan ini mencakup pengumpulan data, pra-pemrosesan untuk memastikan data siap digunakan (meliputi penanganan missing value, penghapusan fitur yang tidak relevan, penanganan outlier, dan standarisasi), serta penerapan metode K-Means Clustering. Pada tahap ini juga dilakukan penentuan jumlah cluster optimal dengan menggunakan metode Elbow**,** Silhouette Coefficient dan DBI**.**
3. ***Objective***: Tujuan dari penerapan metode ini adalah memperoleh segmentasi risiko kredit nasabah yang lebih terukur dan objektif, sehingga dapat membantu bank dalam pengelolaan risiko dan pengambilan keputusan yang lebih tepat.
4. ***Measurement***: Evaluasi dilakukan dengan menghitung Silhouette Score sebagai ukuran kualitas clustering, serta membandingkan hasil pengelompokan untuk mendapatkan jumlah cluster yang paling optimal. Hasil evaluasi ini akan digunakan untuk menentukan model terbaik dan interpretasi profil risiko tiap cluster.