BAB III

METODOLOGI PENELITIAN

3.1 Gambaran Umum Objek Penelitian

Penelitian ini menggunakan data tingkat kesehatan Bank Perekonomian Rakyat (BPR) yang diperoleh dari tim *Data Analytics* OJK untuk periode Juni 2024 hingga Agustus 2025. Dataset berbentuk data *dummy* yang merepresentasikan kondisi keuangan BPR dan terdiri atas 20.340 baris data dengan enam variabel utama, yaitu *CAR, NPL, LDR, BOPO, ROA, dan GCG*. Seluruh variabel tersebut merupakan indikator kerangka RGEC dan digunakan sebagai dasar analisis segmentasi serta prediksi tingkat kesehatan bank. Proses analisis melibatkan dua algoritma *clustering*, *K-Means* dan *K-Medoids*, yang dievaluasi menggunakan *Silhouette Score* dan *Davies–Bouldin Index* untuk menentukan struktur *cluster* terbaik. Hasil klasterisasi tersebut digunakan sebagai label pada tahap klasifikasi dengan algoritma *Random Forest* untuk memprediksi tingkat kesehatan BPR pada data baru. Seluruh tahapan penelitian mengikuti kerangka CRISP-DM, mulai dari pemahaman bisnis, pemahaman data, persiapan data, pemodelan, evaluasi, hingga *deployment* melalui aplikasi web berbasis *Streamlit*.

3.2 Metode Penelitian

Metode penelitian menggunakan pendekatan kuantitatif berbasis machine learning yang berfokus pada proses *clustering* dan klasifikasi untuk mengidentifikasi pola serta karakteristik kesehatan keuangan BPR. Seluruh tahapan analisis dilaksanakan secara sistematis dan terukur untuk menghasilkan model prediktif yang valid, akurat, dan dapat digunakan pada proses pengambilan keputusan.

3.2.1 Perbandingan *Framework*

Dalam data mining, beberapa framework umum digunakan untuk memandu proses analisis, antara lain *Sample, Explore, Modify, Model, Assess* (SEMMA),

Cross-Industry Standard Process for Data Mining (CRISP-DM), dan *Knowledge Discovery in Databases* (KDD). Ketiganya memiliki fokus dan pendekatan yang berbeda, mulai dari eksplorasi data, pemodelan, hingga evaluasi hasil. Untuk mempermudah pemilihan pendekatan yang paling sesuai, Tabel 3.1 merangkum kelebihan, kekurangan, dan fokus utama dari masing-masing framework sebagai dasar pertimbangan dalam penelitian ini.

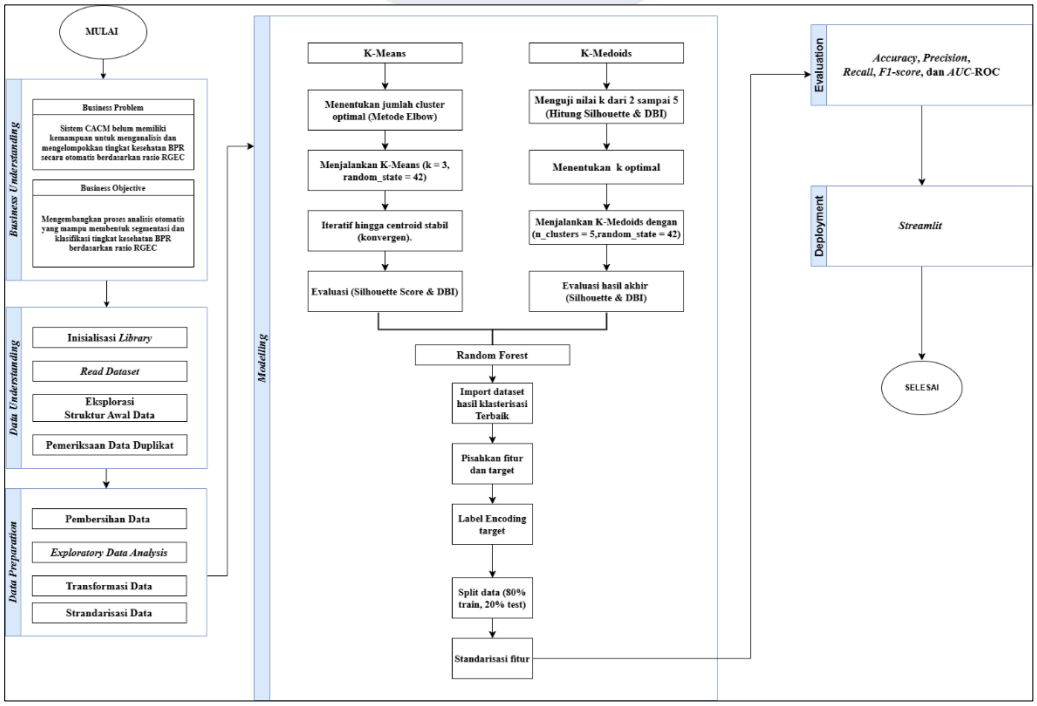
Tabel 3. 1 Perbandingan *Framework*

No.	*Framework*	Kelebihan	Kekurangan	Fokus Utama
1	CRISP-DM [55][56].	• Fleksibel dan dapat diterapkan di berbagai industri dan jenis proyek. • Iteratif dan berfokus pada pemahaman masalah bisnis terlebih dahulu. • Tidak terbatas pada satu jenis perangkat lunak • Memerlukan banyak iterasi, waktu, dan sumber daya untuk validasi hasil.	• Memerlukan banyak iterasi, waktu, dan sumber daya untuk validasi hasil. • Prosesnya bisa menjadi sangat panjang	Pemahaman masalah bisnis dan pengembangan solusi berbasis data.
2	KDD [55][56].	• Fokus pada pemilihan dan persiapan data yang relevan untuk analisis lebih mendalam. • Penekanan pada evaluasi dan interpretasi yang rinci. • Mengakomodasi data yang besar dan kompleks.	• Kurang fokus pada eksplorasi data awal yang mendalam. • Kurang jelas dalam memberikan panduan untuk implementasi hasil	Pembersihan dan transformasi data yang mendalam.
3	SEMMA [55][56].	• Proses yang sangat berfokus pada eksplorasi dan pemodelan data. • Metode yang mudah dipahami dan diterapkan untuk analisis data.	• Terlalu fokus pada proses pemodelan, kurang fleksibel dalam tahap eksplorasi yang lebih mendalam. • Fokus yang lebih sempit pada data tabular, kurang cocok untuk data non-struktural.	Pemodelan data dan eksplorasi dengan pendekatan teknis.

Dari ketiga *framework* yang dibandingkan, CRISP-DM dipilih sebagai *framework* yang paling sesuai untuk penelitian ini. *Framework* ini fleksibel dan dapat mengakomodasi berbagai jenis data, termasuk data rasio keuangan bank seperti *CAR, NPL, BOPO, LDR, GCG, dan ROA.* CRISP-DM menekankan pemahaman masalah bisnis terlebih dahulu, sehingga mendukung analisis yang relevan untuk menilai stabilitas perbankan dan pengambilan keputusan OJK. Keunggulan lainnya adalah pendekatan iteratif yang memfasilitasi optimasi model *clustering* dan prediksi secara berulang, serta kebebasan penggunaan perangkat lunak atau *tools* sesuai kebutuhan analisis, seperti *Python* atau platform *machine learning* lainnya. Dengan proses yang sistematis mulai dari pemahaman data, pemrosesan, pemodelan, hingga evaluasi, CRISP-DM menyokong pengembangan model analisis kesehatan bank secara terstruktur.

3.2.2 Alur Penelitian

Penelitian mengikuti enam tahap CRISP-DM sebagaimana ditunjukkan pada Gambar 3.1. Setiap tahap dijalankan secara iteratif untuk menjaga kualitas data dan memastikan model yang dihasilkan memenuhi tujuan analisis.



Gambar 3. 1 Alur Penelitian

Tahapan CRISP-DM yang digunakan dalam penelitian ini dijelaskan sebagai berikut.

1) Business Understanding

Tahap ini bertujuan untuk merumuskan kebutuhan analitis yang menjadi dasar pengembangan model dalam penelitian. Proses penilaian tingkat kesehatan Bank Perekonomian Rakyat (BPR) di OJK saat ini masih bergantung pada *query* manual melalui *Monitoring Tools* sehingga klasifikasi kesehatan bank belum sepenuhnya terotomatisasi. Kondisi ini menuntut adanya pendekatan berbasis data yang mampu menghasilkan penilaian yang lebih cepat, konsisten, dan objektif. Berdasarkan kebutuhan tersebut, penelitian ini merancang alur analisis menggunakan *machine learning* untuk membentuk segmentasi tingkat kesehatan BPR dan mengembangkan model klasifikasi otomatis. Dua algoritma *clustering K-Means* dan *K-Medoids* digunakan untuk mengidentifikasi pola kesamaan BPR berdasarkan rasio RGEC, sedangkan model *Random Forest* digunakan untuk memprediksi kategori kesehatan dari hasil segmentasi terbaik. Dataset yang dianalisis berupa data dummy yang disusun oleh *tim Data Analytics* OJK untuk periode Juni 2024 hingga Agustus 2025, dengan variabel rasio keuangan yang merepresentasikan kerangka RGEC.

2) Data Understanding

Tahap ini berfokus pada pemahaman karakteristik data yang digunakan dalam analisis. Variabel yang dianalisis mencakup *EntityId, PeriodDate, serta rasio keuangan utama CAR, NPL, BOPO, LDR, ROA, dan GCG* yang mewakili indikator RGEC. Proses yang dilakukan meliputi inisialisasi library, pemanggilan dataset, peninjauan struktur awal data, serta pemeriksaan duplikasi.

3) *Data Preparation*

Tahap *Data Preparation* merupakan proses pembersihan dan transformasi data agar siap digunakan dalam pemodelan. Pada tahap ini, data dibersihkan dari kolom yang tidak relevan serta dilakukan penghapusan duplikasi dan penanganan missing value untuk memastikan integritas dataset. Selanjutnya, dilakukan konversi tipe data agar seluruh variabel berada dalam format numerik yang konsisten. Setelah proses pembersihan, dilakukan analisis deskriptif statistik untuk memahami karakteristik data yang telah bersih dan mendeteksi adanya nilai ekstrem. Data kemudian ditransformasikan ke dalam matriks rasio *Risk Profile, Good Corporate Governance, Earnings, dan Capital* (RGEC) sebagai indikator utama tingkat kesehatan Bank Perekonomian Rakyat (BPR). Tahap akhir adalah standarisasi menggunakan *StandardScaler* agar setiap variabel memiliki skala yang seragam, sehingga hasil analisis *clustering* dan *classification* menjadi lebih akurat dan tidak bias terhadap perbedaan satuan antar variabel.

4) *Modelling*

Tahap *Modelling* berfokus pada penerapan algoritma untuk melakukan segmentasi dan klasifikasi tingkat kesehatan BPR. Dua algoritma *clustering K-Means* dan *K-Medoids* digunakan secara terpisah untuk membentuk kelompok berdasarkan variabel rasio RGEC. Pada *K-Means*, pusat *cluster* diperoleh melalui perhitungan *centroid*, sedangkan *K-Medoids* menggunakan *medoid* sebagai representasi pusat *cluster* sehingga lebih stabil terhadap outlier.

Evaluasi kualitas *cluster* tidak menggunakan akurasi, melainkan *Silhouette Score* dan *Davies–Bouldin Index*, karena kedua metrik ini lebih tepat untuk menilai seberapa baik objek terkelompok dalam *cluster* masing-masing dan seberapa jauh jarak antarkluster. Algoritma dengan kombinasi nilai *Silhouette* tertinggi dan *Davies–Bouldin* terkecil dipilih sebagai struktur *cluster* terbaik.

Hasil klasterisasi tersebut kemudian digunakan sebagai label pada tahap klasifikasi menggunakan *Random Forest.* Dataset dibagi menjadi data latih dan data uji dengan proporsi 80:20. Model *Random Forest* dibangun untuk memprediksi kategori kesehatan BPR pada data baru, dan performanya dievaluasi menggunakan metrik *classification* seperti *accuracy, precision, recall, F1-score, dan ROC-AUC.*

5) Evaluation

Tahap *Evaluation* bertujuan untuk menilai kualitas model dan memastikan tujuan bisnis tercapai. Evaluasi *clustering* dilakukan menggunakan *Silhouette Score* untuk mengukur sejauh mana objek berada dalam *cluster* masing-masing, serta *Davies-Bouldin Index* untuk menilai jarak antar *cluster*, dengan nilai yang lebih kecil menunjukkan kualitas *cluster* lebih baik. Evaluasi classification menggunakan *Accuracy*, *Precision*, *Recall*, *F1-score*, dan *AUC*-ROC untuk mengukur ketepatan prediksi. *Cluster* yang terbentuk merepresentasikan tingkat kesehatan BPR, sedangkan prediksi tingkat kesehatan baru dilakukan secara otomatis dan akurat.

6) Deployment

Tahap *Deployment* merupakan proses penerapan model yang telah dibangun ke dalam sistem berbasis aplikasi web menggunakan *Streamlit*. Aplikasi ini dirancang untuk memudahkan pengguna, khususnya pihak analis atau pengawas dari Otoritas Jasa Keuangan (OJK), dalam melakukan evaluasi tingkat kesehatan Bank Perekonomian Rakyat (BPR) secara cepat dan interaktif. Pada aplikasi ini, pengguna dapat mengunggah data BPR terbaru, dan sistem akan secara otomatis menampilkan hasil prediksi tingkat kesehatan bank secara *real-time* berdasarkan model *Random Forest* yang telah dilatih. Hasil prediksi disajikan dalam tabel interaktif yang berisi kategori tingkat kesehatan masing-masing BPR. Di bawah tabel tersebut, terdapat visualisasi dalam bentuk *barchart* yang menampilkan ringkasan

distribusi kategori kesehatan untuk memberikan gambaran umum kondisi perbankan secara agregat. Selain itu, sistem juga dilengkapi dengan fitur filter berdasarkan kategori kesehatan dan *PeriodDate* agar pengguna dapat menampilkan data sesuai kebutuhan analisis. Seluruh hasil prediksi dan data yang ditampilkan dapat diunduh dalam format Excel maupun CSV, sehingga memudahkan proses dokumentasi dan pelaporan hasil pengawasan.

3.3 Teknik Pengumpulan Data

Data yang digunakan dalam penelitian ini merupakan data dummy yang disediakan oleh tim *data analytic* dari Otoritas Jasa Keuangan (OJK). Data tersebut disusun untuk merepresentasikan kondisi keuangan Bank Perekonomian Rakyat (BPR) berdasarkan rasio keuangan yang mengacu pada pendekatan *Risk Profile, Good Corporate Governance, Earnings, dan Capital* (RGEC). Dataset yang diberikan telah berbentuk file Microsoft Excel (.xlsx) dan berisi rasio keuangan utama seperti NPL, CAR, LDR, ROA, GCG, dan BOPO. Sebelum digunakan, data tersebut diperiksa kembali untuk memastikan kelengkapan, konsistensi, serta kesesuaian format agar dapat diolah menggunakan bahasa pemrograman *Python* pada tahap analisis. Dengan demikian, data yang digunakan telah memenuhi kriteria kelayakan untuk dilakukan pemodelan dan analisis *machine learning* secara akurat dan terstruktur.

3.4 Variabel Penelitian

Penelitian ini menggunakan dua jenis pendekatan analisis, yaitu *clustering* dan *classification*, sehingga variabel penelitian terbagi berdasarkan tahap pemodelan yang dilakukan :

3.4.1 Variabel Independen

Variabel independen dalam penelitian ini terdiri atas rasio-rasio keuangan yang berfungsi sebagai indikator dalam kerangka penilaian RGEC. Seluruh rasio tersebut telah ditransformasikan ke dalam bentuk skala penilaian (Peringkat Kesehatan atau PK) dengan rentang nilai 1 hingga 5 sesuai dengan kriteria

penilaian yang ditetapkan oleh Otoritas Jasa Keuangan (OJK). Adapun variabel yang digunakan meliputi rasio kredit bermasalah (PK_NPL), rasio kecukupan modal (PK_CAR), rasio likuiditas pinjaman terhadap simpanan (PK_LDR), rasio profitabilitas (PK_ROA), skor tata kelola perusahaan yang baik (PK_GCG), serta rasio efisiensi operasional (PK_BOPO). Keenam variabel tersebut digunakan sebagai input *features* pada tahap klasterisasi menggunakan algoritma *K-Means* dan *K-Medoids*, serta pada tahap klasifikasi menggunakan algoritma *Random Forest* untuk membangun model prediksi tingkat kesehatan BPR.

3.4.2 Variabel Dependen

Variabel dependen dalam penelitian ini berbeda berdasarkan tahap analisis yang dilakukan, yaitu:

1) Pada tahap *clustering*, variabel dependen direpresentasikan oleh hasil pengelompokan (*cluster* label) yang menunjukkan kategori kelompok BPR berdasarkan kesamaan pola rasio keuangan. Setiap label *cluster* merepresentasikan kelompok BPR dengan kondisi keuangan yang serupa, misalnya *Cluster* 0 dikategorikan sebagai Sangat Sehat dan *Cluster* 1 sebagai Kurang Sehat
2) Pada tahap *classification*, variabel dependen berupa kategori tingkat kesehatan BPR yang digunakan sebagai target dalam model *Random Forest*. Kategori ini diperoleh dari hasil klasterisasi terbaik (*K-Medoids*) dan dibagi ke dalam lima kelas, yaitu Sangat Sehat, Sehat, Cukup Sehat, Kurang Sehat, dan Tidak Sehat.

3.5 Teknik Analisis Data

Proses analisis data dimulai dengan pengumpulan rasio keuangan berbasis RGEC yang disimpan dalam format *Excel* dan diolah menggunakan pandas untuk pembersihan data serta *NumPy* untuk manipulasi numerik. Tahap persiapan data mencakup pemeriksaan duplikasi, penanganan nilai hilang, konversi tipe data, serta standardisasi menggunakan *StandardScaler* agar seluruh fitur berada pada skala yang seragam. Eksplorasi awal dilakukan melalui visualisasi menggunakan *matplotlib* dan

seaborn untuk memahami pola sebaran dan mendeteksi potensi *outlier*. Setelah data siap, proses *clustering* dilakukan menggunakan algoritma *K-Means* dan *K-Medoids* melalui *library scikit-learn* dan *scikit-learn-extra*. Kualitas pembentukan *cluster* dievaluasi menggunakan *Silhouette Score* dan *Davies Bouldin Index* untuk menentukan struktur *cluster* yang paling representatif. Hasil klasterisasi terbaik kemudian dijadikan label awal pada tahap klasifikasi.

Dataset dibagi menjadi data latih dan data uji dengan proporsi 80:20 menggunakan *train_test_split*. Tahap pemodelan kemudian dilanjutkan dengan pembangunan model *Random Forest* melalui *RandomForestClassifier*, yang dipilih karena stabil, mampu menangani data multivariat, serta minim *overfitting*. Evaluasi performa dilakukan menggunakan metrik *accuracy, precision, recall, F1-score, ROC-AUC,* serta analisis melalui *confusion matrix* dan *classification_report.* Tahap akhir berupa deployment model ke dalam aplikasi web berbasis *Streamlit*. Model yang telah dilatih disimpan menggunakan joblib dan dimuat ulang di aplikasi untuk menghasilkan prediksi tingkat kesehatan BPR secara *real-time*. Aplikasi ini menampilkan tabel hasil prediksi dan visualisasi pendukung, sehingga dapat digunakan oleh analis sebagai alat bantu pengawasan yang lebih cepat dan praktis.