BAB II
TINJAUN PUSTAKA

1.1 Penelitian Terkait

Penelitian sebelumnya yang menjadi latar belakang pada penelitian ini, dijabarkan pada tabel dibawah ini :

Tabel 2.1 Tabel Penelitian Terkait

No.	Judul, penulis, tahun	Dataset	Metode	Hasil
1.	Klasifikasi Data Mining untuk Prediksi Potensi Nasabah dalam Membuat Deposito Berjangka Nia Nuraeni 2021	dataset pada penelitian ini didapatkan dari repositori UCI dengan alamat url *https://archive.ics.uci.edu/ml/datasets/Bank+Marketing* dengan jumlah record sebanyak 41188 record. Dari 41188 Record yang didapatkan dari Repository UCI, 4640 nasabah membuat deposito berjangka sedangkan sisanya sebanyak 36548 tidak membuat deposito berjangka	Decision Tree (C4.5), Naïve Bayes dan k-Nearest Neighbor.	Hasil penelitian prediksi potensi nasabah yang membuka deposito berjangka dengan algoritma klasifikasi Decision Tree, Naïve Bayes Dan K-Nearest Neighbor maka ditetapkan bahwa akurasi tertinggi pada algoritma klasifikasi adalah algoritma decision tree dengan hasil akurasi 91.26%
2.	Analisis Data Bank Direct Marketing dengan Perbandingan	dataset pada penelitian ini didapatkan dari repositori UCI dengan alamat url (https://archive.ics.uci.e	algoritma Naïve Bayes, K-	Pada penelitian ini telah dilakukan penghitungan metode Naive Bayes, K-Nearest

	Klasifikasi Data Mining Berbasis Optimize Selection (Evolutionary) Ahmad Fauzi 2021	du/ml/datasets/Bank+Marketing). Dari data yang diambil penulis memilih dataset yang berjumlah 4.521, dengan 17 atribut di mana dari 17 atribut tersebut terdapat satu atribut prediksi	Nearest Neighbor, Support Vector Machine dengan optimasi Optimize Selection (Evolutionary)	Neighbor, dan Support Vector Machine, dengan metode optimasi Optimize Selection (Evolutionary), dari dataset Bank Direct Marketing. a. Algoritma Naive Bayes, K-Nearest Neighbor, dan Support Vector Machine, dapat di implementasikan dalam pengklasifikasian dataset Bank Direct Marketing pada aplikasi rapidminer 5.3. b. Dengan penambahan metode optimasi Optimize Selection (Evolutionari) mampu meningkatkan hasil dari nilai akurasi pada penghitungan algoritma Naive Bayes, K-Nearest Neighbor, dan

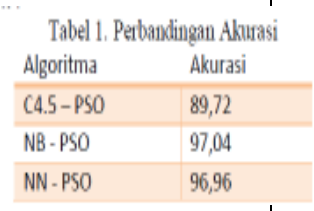
				Support Vector Machine. Nilai akurasi tertinggi adalah 90.18% dari hasil penghitungan algoritma Naive Bayes. c. Dalam peningkatan nialai akurasi dari algoritma Naive Bayes, K-Nearest Neighbor, dan Support Vector Machine, dengan metode optimasi menghasilkan nilai akurasi yang berbeda yaitu algoritma Naive Bayes 3.47%, K-Nearest Neighbor 2.28%, dan Support Vector Machine menghasilkan peningkatan nilai akurasi tertinggi yaitu 8.47%.
3.	Analisis Prediksi Nasabah yang	Data yang digunakan dalam penelitian ini berasal dari data nasabah marketing	Algoritma C4.5 dan	Hasil teknik klasifikasi dengan algoritma C4.5 pada penelitian ini

	Berpotensi Membuka Deposito pada Bank Umum di Bekasi Menggunakan Algoritma C4.5 dan Naive Bayes Dwi Handayani dan Nuryuliani 2020	suatu Bank Umum di Bekasi pada tahun 2018. Data yang diuji menggunakan tujuh atribut sebagai predictor yang terdiri dari Pekerjaan, Status, Pendidikan, KPR, Loan, Kontak, Pemasaran_ Sebelumnya, dan dan satu atribut sebagai penentu predictor yaitu Pembukaan_Deposito yang berjumlah 9374 record data dengan komposisi data training 80% berjumlah 7499 record dan data testing 20% berjumlah 11875 record	Naive Bayes	menghasilkan persentase keakuratan prediksi sangat baik pada nasabah yang berpotensi membuka deposito dengan nilai accuracy berjumlah 91,9467%, precision berjumlah 92,40%, recall berjumlah 91,90%, dan f-measure berjumlah 92% dibandingkan dengan algoritma Naive Bayes dengan nilai accu-racy berjumlah 89,8133%, precision berjumlah 90,2%, recall berjumlah 89,8%, dan f-measure berjumlah 89,8% . Melihat nilai accuracy hampir mendekati 100%, maka teknik klasifikasi dengan algoritma C4.5

				merupakan kategori klasifikasi sangat baik digunakan untuk memprediksi nasabah yang berpotensi membuka deposito.
4.	Klasifikasi Pelanggan Deposito Potensial menggunakan Ensembel Least Square Support Vector Machine 1Firman Aziz, 2Jeffry 2020	Data yang diolah adalah data bank direct marketing dataset yang dapat diakses melalui University of California at Irvine (UCI) Machine Learning Repository. Data tersebut berasal dari bank Portugal, dari bulan mei 2008 sampai bulan juni 2013. Dari total jumlah data keseluruhan sebanyak 41.188 pelanggan, dinormalisasi menjadi 9.280 untuk menyeimbangkan anatar pelanggan deposito berjangka dan bukan pelanggan deposito berjangka. Jadi data yang diolah sebanyak 4,640 data pelanggan deposito berjangka dan 4,640	peneliti an ini mengus ulkan pengem bangan dari metode support vector machin e yaitu metode least square support vector machin e kemudi an di ensemb le menggu nakan	Hasil menunjukkan bahwa metode yang diusulkan yakni ensemble least square support vector machine lebih baik dibandingkan dengan metode lainnya dengan persentase tingkat accuracy, sensitivity, specivicity masing-masing adalah 95.15%, 92.93%, 97.61% dengan total rata-rata hasil klasifikasi sebesar 95.23%.

		data bukan pelanggan deposito berjangka	boostin g	
5.	Perbandingan Metode Klasifikasi Data Mining untuk Nasabah Bank Telemarketing Pungkas Subarkah, Enggar Pri Pambudi, Septi Oktaviani Nur Hidayah 2020	Dataset diambil dari database UCI Repository. Dataset terdiri dari 4521 record, memiliki 17 attribute (16 attribute dan 1 attribute target),	algorit ma Classifi cation and Regress ion Trees (CART) dan naive bayes	hasil penelitian ini, peneliti mengambil kesimpulan bahwa hasil klasifikasi algoritma CART memiliki nilai akurasi sebesar 89.51% lebih baik daripada algoritma naive bayes yaitu dengan nilai akurasi sebesar 86.88%., pada dataset bank telemarketing, dengan nilai selisih 2.63%.
6.	Prediksi Nasabah Yang Berpotensi Membuka Simpanan Deposito Menggunakan Naive Bayes Berbasis Particle Swarm Optimization Alvino Dwi Rachman	penulis menggunakan dataset public berjudul Bank Marketing yang pernah digunakan oleh S.Moro, R. Laureano dan P. Cortez, yang berjudul Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology dari website University of	algorit ma Naive Bayes berbasi s PSO	algoritma Particle Swarm Optimization (PSO) ternyata dapat mengoptimalkan algoritma Naive Baiyes dengan pembobotan atribut terlebih dahulu sehingga akurasi dapat meningkat. Terbukti, pada dataset bank

	Prabowo1, Muljono2 2018	California, Irvine (UCI) Machine Learning. Dataset Bank Marketing tersebut terdiri dari 16 attribut prediktor dan 1 attribut label		marketing, akurasi algoritma Naive Baiyes adalah 82,19% dan setelah menggunakan algoritma Particle Swarm Optimization untuk mengoptimalkan algoritma Naive Baiyes dengan pembobotan atribut, akurasi meningkat menjadi 89,70%, yang berarti akurasi meningkat 7,51%.
7.	Analisis Perbandingan Algoritma Optimasi pada Random Forest untuk Klasifikasi Data Bank Marketing Yoga Religia1, Agung Nugroho2, Wahyu Hadikristanto3	Data yang digunakan dalam penelitian ini diambil dari situs www.data.world	menggunakan algoritma RF	Berdasarkan pengujian yang telah dilakukan diperoleh hasil bahwa performa paling optimal dari klasifikasi data Bank Marketing adalah dengan menggunakan algoritma RF dengan akurasi sebesar 88,30%, AUC (+) sebesar

	2021			0,500 dan AUC(-) sebesar 0,000
8.	Analisa Nasabah Potensial Tabungan Deposito Berjangka Menggunakan Teknik Klasifikasi Data Mining Candra Agustina 2018	Data yang digunakan dalam penelitian ini diambil dari link https://archive.ics.uci.edu/ml/datasets/bank+marketing yang terdiri dari 17 atribut.	Decision Tree, Naïve Bayes, Algoritma C.4.5	Hasil pengolahan keseluruhan dapat dilihat pada tabel Berikut :  Tabel 1. Perbandingan Akurasi Algoritma | Akurasi C4.5 – PSO | 89,72 NB - PSO | 97,04 NN - PSO | 96,96 Dengan demikian Naïve Bayes dengan Particle Swarm Optimization memiliki akurasi paling tinggi yaitu 97,04

Dari hasil beberapa review jurnal nasional dan internasional yang saya baca, saya menyimpulkan bahwa agar dataset dapat terakurasi dengan baik maka di butuhkan suatu pengembangkan model algoritma untuk mengklasifikasi calon nasabah bank pada produk deposito agar mendapatkan akurasi tertinggi. dan metode yang digunakan untuk beberapa jurnal yang saya review tingkat akurasi cenderung lebih tinggi dengan menggunakan algoritma Naïve Bayes dan Particle Swarm Optimization (PSO).

Hasil dapat dilihat pada tabel berikut :

Tabel 2.2 Tabel Hasil Akurasi Review Jurnal

No. Jurnal	Metode Terbaik dan akurasi
1.	decision tree akurasi 91.26%, menggunakan tools Rapid Miner 9.8.
2.	Naive Bayes akurasi 90.18%, menggunakan tools Rapidminer 5.3.
3.	Algoritma C4.5 akurasi 91,9467%, precision 92,40%, recall 91,90%, dan f-measure 92% menggunakan tools Weka
4.	support vector machine + AdaBoost akurasi 95.23%.
5.	algoritma CART akurasi 89.51% menggunakan tools Weka
6.	Naive Baiyes akurasi 82,19%, dan setelah menggunakan algoritma Particle Swarm Optimization akurasi meningkat menjadi 89,70%, yang berarti akurasi meningkat 7,51%.
7.	algoritma RF akurasi 88,30%, menggunakan model perbandingan antara penggunaan Random Forest (RF) dan RF dengan optimasi Bagging dan Genetic Algoritm (GA)
8.	Naïve Bayes + Particle Swarm Optimization akurasi 97,04, penenlitian ini menghilangkan atribut karena mempunyai bobot kurang dari 0,500.

1.2 Data Mining

Definisi data mining adalah proses untuk menemukan dan memberikan gambaran berupa pola struktural dalam data sehingga dapat digunakan sebagai alat bantu untuk membantu memberikan penjelasan terhadap sekumpulan data, selain itu juga membuat prediksi dari data tersebut(Agustina 2018).

Data mining salah satu cara untuk memunculkan pengetahuan yang ada di database atau Knowledge Discovery in Database (KDD). Untuk mengetahui pola dan model suatu data perlu di lakukan ekstraksi data yang berguna untuk melihat informasi yang ada di dalamnya. Dengan adanya hubungan KDD dan

data mining menyebabkan korelasi antar keduanya sangat dibutuhkan (Darmawan et al. 2018).

Data mining merupakan bagian dari tahapan proses Knowledge Discovery in Database (KDD). Dengan data mining, kita dapat melakukan proses klasifikasi, prediksi, perkiraan dan mendapatkan informasi lain yang bermanfaat dari kumpulan data dengan jumlah besar. Data mining adalah proses pencarian pola tetrtentu atau informasi yang menarik dalam data yang sudah dipilih dengan menggunakan teknik atau metode tertentu. teknik, metode maupun algoritma dalam data mining sangat bervariasi. Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses Knowledge Discovery in Database secara keseluruhan. Algoritma Naïve Bayes, k-Nearest Neighbor dan Decision Tree merupakan contoh penerapan metode klasifikasi pada data mining(Nuraeni 2021).

1.3 Rapidminner

Rapidminer merupakan perangkat lunak yang digunakan untuk pengolahan data. Dengan menggunakan prinsip dan algoritma data mining. Rapidminer mengekstrak pola-pola dari dataset yang besar dengan mengkombinasikan metode statistika, kecerdasan buatan, dan database. Rapidminer memudahkan penggunanya dalam melakukan perhitungan data yang sangat banyak menggunakan operator-operator. Operator ini berfungsi untuk memodifikasi data. Data dihubungkan dengan node-node pada operator kemudian kita hanya tinggal melihat hasilnya. Hasil yang diperlihatkan dapat ditampilkan secara visual menggunakan grafik. Hal ini yang menjadikan rapidminer sebagai software pilihan untuk melakukan ekstraksi data dengan metode-metode data mining(Hidayati and Nugroho 2021).

1.4 Klasifikasi

Klasifikasi merupakan bagian dari prediksi, dimana nilai yang diprediksi berupa label. Klasifikasi menentukan class atau grup untuk tiap contoh data, input dari model klasifikasi adalah atribut dari contoh data (data samples) dan outputnya adalah class dari data samples itu sendiri, dalam machine learning

untuk membangun model klasifikasi digunakan metode supervised learning. Metode supervised learning yaitu metode yang mencoba untuk menemukan hubungan antara atribut masukan dan atribut target, hubungan yang ditemukan diwakili dalam struktur yang disebut model(Mandiri 2016).

Dalam klasifikasi kita dapat menentukan orang atau objek kedalam suatu kategori tertentu, contoh untuk masalah klasifikasi adalah menentukan apakah seseorang pasien “mengidap” atau “tidak mengidap” penyakit tertentu. Informasi tentang pasien sebelumnya digunakan sebagai bahan untuk melatih algoritma untuk mendapatkan rule atau aturan(Mandiri 2016).

Salah satu tujuan klasifikasi adalah untuk meningkatkan kehandalan hasil yang diperoleh dari data(Mandiri 2016).

1.5 Decision Tree

Decision Tree merupakan alat pendukung keputusan yang menggunakan grafik atau model seperti pohon untuk mendeskripsikan hubungan antara variable yang berbeda. Decision Tree atau pohon keputusan dihasilkan dengan cara membagi setiap nilai dari atribut menjadi cabang dalam setiap kemungkinan. Cara kerjanya adalah dengan menelusuri dari akar sampai ke cabang hingga mencapai class suatu object ditemukan. Salah satu algoritma yang dapat digunakan untuk membuat pohon keputusan adalah algoritma C4.5. Algoritma C4.5 merupakan algoritma pengembangan dari Algoritma ID3 sebelumnya yang diciptakan oleh J. Rose Quinlan(Nuraeni 2021).

Secara umum algoritma C4.5 dalam membangun pohon keputusan (Decision Tree) sebagai berikut(Nuraeni 2021):

a. Pilih atribut sebagai akar
b. Buat cabang untuk tiap-tiap nilai
c. Bagi case dalam setiap cabang

Ulangi proses dalam setiap cabang sampai semua case pada cabang memiliki class yang sama.

1.6 Split Validation

Validation adalah teknik validasi yang membagi data menjadi dua bagian secara acak, sebagian sebagai data training dan sebagian lainnya sebagai data testing(Subarkah et al. 2020). Dengan menggunakan Split Validation akan dilakukan percobaan training berdasarkan split ratio yang telah ditentukan sebelumnya, untuk kemudian sisa dari split ratio data training akan dianggap sebagai data testing. Data training adalah data yang akan dipakai dalam melakukan pembelajaran sedangkan data testing adalah data yang belum pernah dipakai sebagai pembelajaran dan akan berfungsi sebagai data pengujian kebenaran atau keakurasian hasil pembelajaran(Aziz 2020).

1.7 Seleksi Fitur

Seleksi fitur merupakan proses yang melibatkan subset dari kumpulan fitur yang menghasilkan keluaran seperti keseluruhan kumpulan fitur. Seleksi fitur biasanya digunakan untuk memilih fitur yang optimal, mereduksi dimensi, meningkatkan akurasi algoritma klasifier, dan menghapus fitur yang tidak relevan(Rahmawati et al. 2020). Tujuan utama dari seleksi fitur adalah untuk mengurangi jumlah fitur yang digunakan dalam klasifikasi dengan tetap menjaga akurasi klasifikasi yang dapat diterima. Pemilihan fitur dapat berdampak besar pada keefektifan algoritma klasifikasi yang dihasilkan, dalam beberapa kasus, sebagai hasil dari pemilihan fitur, akurasi klasifikasi yang akan datang dapat ditingkatkan(Puteri and Achmad 2021).

Manfaat melakukan pemilihan fitur sebelum memodelkan data Anda adalah sebagai berikut:

a. Mengurangi Overfitting: Data yang lebih sedikit berarti lebih sedikit kesempatan untuk membuat keputusan berdasarkan noise.
b. Meningkatkan Akurasi: Data yang kurang menyesatkan berarti akurasi pemodelan meningkat.
c. Mengurangi Kompleksitas: lebih sedikit titik data mengurangi kompleksitas algoritme dan membuatnya lebih mudah dipahami.
d. Pelatihan Lebih Cepat: Ini memungkinkan algoritme pembelajaran mesin untuk berlatih lebih cepat.

e. Pada sistem ini digunakan dua proses seleksi fitur yaitu proses seleksi fitur sekuensial/forward dan proses seleksi fitur mundur. Pada sistem ini digunakan dua proses seleksi fitur yaitu proses seleksi fitur sekuensial/forward dan proses seleksi fitur mundur.

1.8 Particle Swarm Optimization (PSO)

PSO adalah metode pencarian populasi, yang berasal dari penelitian untuk pergerakan berkelompok burung dan ikan dalam mencari makan. Seperti algoritma genetika, PSO melakukan pencarian menggunakan populasi (disebut swarm) dari individu-individu (disebut partikel) yang diperbarui dari setiap iterasi yang dilakukan. Untuk menemukan solusi yang optimal, setiap partikel bergerak ke arah posisi terbaik yang sebelumnya (pbest) dan posisi global terbaik (gbest)(Dwi et al. 2018).

Pada penelitian ini, PSO digunakan untuk menyeleksi fitur (*feature selection*). Untuk menyeleksi fitur atau atribut tersebut, maka PSO menggunakan bobot atribut yang telah dihitung dan atribut yang telah diseleksi akan diprediksi menggunakan algoritma *Decision Tree*. Langkah-langkah untuk mendapatkan bobot tersebut adalah sebagai berikut(Dwi et al. 2018).

1. Menyiapkan sampel dataset, siap untuk menghitung bobot dari record pertama.
2. Inilisialisasi populasi (swarm) P(t) sehingga ketika t = 0, lokasi $xi(t)$ dari setiap partikel $Pi \in P(t)$ di suatu ruang yang luas menjadi acak.
3. Melalui setiap posisi partikel saat itu, $xi(t)$ mengevaluasi performa dari F.
4. Membandingkan kinerja masing-masing individu saat ini dengan individu yang mempunyai kinerja terbaik yang sejauh ini dimiliki jika $F(xi(t)) > pbesti$ maka

5. Membandingkan kinerja masing-masing partikel dengan kinerja partikel terbaik secara global, jika $F(xi(t)) > gbesti$ maka

$$\begin{cases} gbest_i = F(x_i(t)) \\ X_{gbest_i} = x_i(t) \end{cases}$$

Ubah kecepatan vector dari particle dengan menggunakan formula 5. Setelah mendapatkan kecepatannya lalu, tempatkan setiap partikel ke lokasi yang baru menggunakan formula 6. Kembali ke langkah nomor 2, ulangi perulangan hingga mendapatkan nilai yang konvergen dan mendapatkan bobot.

6. Jika perhitungan bobot dari record terakhir di data sampel selesai, maka akhiri. Jika tidak, siap untuk menghitung bobot dari record berikutnya dan kembali mengulangi dari langkah nomor 1.
7. Setelah itu, lalu hitung rata-rata bobot yang telah dihitung dari data sampel tersebut.

1.9 Algoritma C4.5

Algoritma C4.5 yaitu algoritma yang digunakan untuk membentuk sebuah pohon keputusan. Didalam algoritma ini, pohon-pohon keputusan yang dibuat berdasarkan parameter-paramater yang dapat membangun keputusan(Yulianti et al. 2020).

Algoritma C4.5 merupakan pengembangan dari algoritma konvensional induksi pohon keputusan yaitu ID3. Algoritma yang merupakan pengembangan dari ID3 ini dapat mengklasifikasikan data dengan metode pohon keputusan yang memiliki beberapa kelebihan. Adapun kelebihannya dapat mengolah data numerik (kontinyu) dan diskret, dapat menangani nilai atribut yang hilang, menghasilkan aturan-aturan yang mudah diinterpretasikan, dan tercepat diantara algoritma algoritma yang menggunakan memori utama di komputer(Umum et al. 2020).

1.10 Akurasi

Accuracy merupakan hasil dari klasifikasi yang memiliki label dengan tipe atribut polynominal. Penjelasan dari accuracy adalah angka yang menunjukkan seberapa dekat nilai yang terukur (measured) dengan nilai sebenarnya (reality). Fungsi dari accuracy ialah untuk mengungkapkan kebenaran suatu pengukuran

dan ditentukan secara absolut dan komparatif (Navisa et al. 2021). Perhitungan accuracy dihitung dengan mengambil persentase prediksi yang benar yang di atas jumlah dengan rincian sebagai berikut(Navisa et al. 2021):

$$Accuracy = \frac{Sum\ of\ true\ positives + Sum\ of\ true\ negatives}{Total\ population}$$

1.11 Confusion Matrix

Penggunaan teknik Confusion Matrix biasanya digunakan dalam penelitian untuk mengevaluasi tingkat akurasi yang dihasilkan dari teknik klasifikasi pada proses mining. Penggunaan Confusion Matrix juga digunakan untuk mengetahu apakah klasifikasi tersebut bersifat positif atau negatif serta adakah kesalahan dalam melakukan proses klasifikasi (Darmawan et al. 2018).

1.12 Kurva ROC

Kurva ROC banyak digunakan untuk menilai hasil prediksi, kurva ROC adalah teknik untuk memvisualisasikan, mengatur, dan memilih pengklasifikasian berdasarkan kinerja mereka. Kurva ROC adalah tool dua dimensi yang digunakan untuk menilai kinerja klasifikasi yang menggunakan dua class keputusan, masingmasing objek dipetakan ke salah satu elemen dari himpunan pasangan, positif atau negatif. Pada kurva ROC, TP rate diplot pada sumbu Y dan FP rate diplot pada sumbu X(Mandiri 2016).

Untuk klasifikasi data mining, nilai AUC dapat dibagi menjadi beberapa kelompok.

a) 0.90-1.00 = Excellent Classification
b) 0.80-0.90 = Good Classification
c) 0.70-0.80 = Fair Classification
d) 0.60-0.70 = Poor Classification
e) 0.50-0.60=Failur

The Area Under Curve (AUC) dihitung untuk mengukur perbedaan performasi metode yang digunakan. AUC dihitung menggunakan rumus:

$$\theta^r = \frac{1}{mn}\sum_j^n = 1\sum_i^m = 1\,\psi(xi^r, xj^r$$

Dimana

$$g(X,Y) = \begin{cases} 1 & Y < X \\ \frac{1}{2} & Y = X \\ 0 & Y > X \end{cases}$$

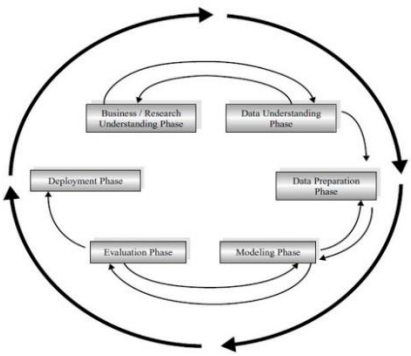
X= Output Positif

X= Output Positif

Y = Output Negatif

1.13 Metode CRISP-DM

CRISP menyediakan proses standar nonproprietary dan tersedia secara bebas untuk memasukkan data mining ke dalam strategi pemecahan masalah umum dari bisnis atau unit penelitian. Menurut CRISP-DM, proyek data mining tertentu memiliki siklus hidup yang terdiri dari enam fase, seperti yang diilustrasikan pada Gambar 2.5(Larose n.d.). Perhatikan bahwa urutan fase adaptif. Artinya, fase berikutnya dalam urutan sering tergantung pada hasil yang terkaitdengan fase sebelumnya. Ketergantungan paling signifikan antara fase ditunjukkan oleh panah. Sebagai contoh, anggaplah kita berada dalam fase pemodelan. Tergantung pada perilaku dan karakteristik model, kita mungkin harus kembali ke fase persiapan data untuk penyempurnaan lebih lanjut sebelum melanjutkan kemodel fase evaluasi.Sifat iteratif CRISP dilambangkan dengan lingkaran luar pada Gambar 2.3(Larose n.d.).



Gambar 2.1 Tahapan Metode CRISP-DM

Seringkali, solusi untuk masalah bisnis atau penelitian tertentu mengarah ke pertanyaan menarik lebih lanjut, yang kemudian dapat diserang menggunakan proses umum yang sama seperti sebelumnya. Pelajaran dari proyek-proyek

masa lalu harus selalu dibawa sebagai masukan ke dalam proyek-proyek baru. Berikut ini adalah garis besar dari setiap fase. Meskipun mungkin, masalah yang dihadapi selama fase evaluasi dapat mengirim analis kembali ke salah satu fase sebelumnya untuk perbaikan, untuk kesederhanaan kami hanya menunjukkan loop yang paling umum,kembali ke tahap pemodelan(Larose n.d.).