BAB II

TINJAUAN PUSTAKA

2.1 Penelitian terkait

Penelitian ini menggunakan referensi yang diambil dari berbagai tulisan, baik dari buku, jurnal penelitian nasional maupun internasional serta jurnal prosiding sebagai bahan untuk menyusun teori pendukung. Beberapa jurnal dan artikel yang memiliki keterkaitan dengan topik penelitian ini diantaranya:

1. Perbandingan 4 algoritma berbasis particle swarm optimization (PSO) untuk memprediksi kelulusan tepat waktu mahasiswa. (Zainuddin, 2019)

   Penelitian tersebut bertujuan untuk mencari algoritma terbaik diantara naïve bayes, decision tree, k-nearest neighbor dan neural network yang berbasis PSO. Hasil menunjukkan algoritma KNN berbasis PSO mempunyai performa terbaik dengan akurasi 74.08% dan nilai AUC 0.788. disimpulkan bahwa fitur PSO dapat meningkatkan nilai akurasi suatu algoritma.
2. Optimasi penjadwalan mata kuliah di jurusan teknik informatika PENS dengan menggunakan algoritma particle swarm optimization (PSO). (Ariani, Fahriza, & Prasetyaningrum, 2011)

   Jurnal tersebut membahas tentang membuat penjadwalan otomatis agar tidak terjadi bentrok jadwal. Pemilihan PSO untuk optimasi karena algoritma ini memiliki tool-tools yang bisa membantu dalam hal optimasi. Parameter yang digunakan yakni C1=1.5, C2=1.5, W=0.5 dan jumlah partikel sebanyak 10.

3. An analysis of particle swarm optimization with data clustering-tecnique for optimization in data mining. (Khan, Bawane, & Bodkhe, 2010)

   Data klaster merupakan pendekatan salam hal menemukan kelas, konsep atau grup dalam suatu pola. Klastering merepresentasikan untuk kebutuhan data yang besar. PSO dijadikan fokus penelitian untuk menangani pengklasteran tersebut. Hasilnya PSO efektif dalam pengoptimasian global untuk permasalahn variabel.

4. Perbandingan performansi algoritma cross entropy (CE) dan algoritma particle swarm optimization (PSO) pada penyelesaian permasalahan flowshop scheduling. (Krisnawati, 2011)

   Penjadwalan flowshop yaitu penjadwalan proses produksi dari n-job yang memiliki urutan proses produksi yang sama. Algoritma yang dipakai yakni cross entropy(CE) dan PSO. Cross Entropy(CE) dinilai menghasilkan solusi penurunan jumlah sampel memberikan hasil relatif lebih baik dari PSO, sedangkan waktu komputasi, PSO lebih singkat dibanding CE.

5. Komparasi dan analisis kinerja model algoritma SVM dan PSO-SVM. (Sasongko, 2016).Menggunakan seleksi atribut dalam prosesnya menggunakan algoritma PSO dalam SVM. Dan SVM sendiri. Dilakukan dengan dataset peminatan siswa SMA ABC yang melibatkan 280 siswa dengan menggunakan beberapa kernel yakni dot, radial, polynominal, neural dan anova kernel. Hasilnya PSO dapat meningkatkan akurasi, presisi, recall, AUC dari model SVM. Akurasi SVM-PSO pada engujian anova sebesar 99.30%. Performansi yang diimplementasikan mendapatkan akurasi sebesar 99.29%.

6. Algoritma klasifikasi data mining naïve bayes berbasis particle swarm optimization untuk mendeteksi penyakit jantung. (Nur Aeni Widiastuti:2014)

   Dataset yang digunakan adalah data hasil laboratorium dan hasil rekam jantung yang terdiri atas usia, jenis kelamin, hasil uji darah, thorax, hasil rekam jantung, dll. Hasil akurasi naïve bayes sebesar 82.14% dengan nilai AUC 0.686 dengan kategori poor clasification. Pada experimen kedua yang melibatkan PSO untuk optimasinya menghasilkan 92.86% dengan nilai AUC 0.839 dengan kategori good clasiification.

7. Klasifikasi pemanfaatan programberas sejahtera berdasarkan tingka kemiskinan menggunakan decision tree C4.5 berbasis particle swarm optimization. (Waluyo, 2017).

   Penulis mengangkat masalah tidak tepat nya penyaluran bantuan beras yang dibagikan karena terkadang terdapat warga yang mampu tapi masuk dalam kategori tidak mampu sehingga mendapat jatah pembagian beras. Penulis memilih algoritma C45 karena dapat membentuk kriteria yang prioritas ditambah dengan algoritma PSO untuk meingkatkan akurasi atribut prioritas tadi. Atribut yan digunakan yakni pekerjaan, pendapatan, listrik, lantai, dinding, sumber air dan tv. Dengan kombinasi C4.5 berbasis PSO dihasilkan akurasi sebesar 97.67%

8. Perbandingan aliran daya optima mempertimbangkan biaya pembangkitan dan kestabilan daya menggunakan particle swarm optimization dan algoritma genetika.(Muhammad Saukani, Ermanu Azizul H., 2016)

   Analisis Optimal Power Flow(OPF) untuk mengetahui besar daya efektif yang harus dibangkitkan pada setiap pembangkit listrik sistem tenaga listrik. Tujuannya

untuk menekan biaya operasional pembangkitan dan mengurangi daya pada proses distribusi. Penelitian tersebut menggunakan metode newton rapshon dan optimasi PSO dan AG. Hasil simulasi menunjukkan bahwa PSO dan AG dapat melakukan penghematan biaya masing-masing 3.08% dan 7.18% dibandingkan metode newton rapshon, dengan metode AG 220% lebih efektif dari PSO. Hasil optimasi terbaik menggunakan AG lebih unggul dari PSO pada sisi biaya pembangkitan sedangkan untuk daya pembangkitan dan sisi profil tegangan metode PSO lebih baik.

9. Analisis komparasi pemodelan algoritma decision tree menggunakan metode particle swarm optimization dan adabost untuk prediksi awal penyakit jantung.(Jusia, 2018)

   Penelitian tersebt melakukan improve clasification accuracy dengan memodifikasi pemodelan algoritma klasifikasi decision tree yang ditambahkan dengan adabost dan particle swarm optimization. Dataset yang digunakan hasil extraksi dari [ublic dataset yang diambil dari arsip university of california irvine sebanyak 270 record. Hasil evaluasi bahwa decision tree punya nilai 79.26% dan nilai AUC 0.889. setelah dilakukan modifikasi dengan algoritma PSO menghasilkan nilai akurasi 82.59% dan nilai AUC 0.916. sedangkan adabost bernilai akurasi 79.26% dan AUC 0.955.

10. Perbandingan algoritma genetika dan particle swarm optimization dala optimasi penjadwalan matakuliah.(Marbun, Nikentari, & Bettiza, 2013)

Dalam penelitian ini dibangun aplikasi untuk menyelesaikan masalah penjadwalan dengan membandingkan 2 algoritma optimasi yaitu genetika dan PSO. GA mampu menyelesaikan permasalahan penjadwalan matakuliah dengan jumlah 42 matakuliah iterasi ke 10 dalam waktu 8.79 detik dengan fitness terbaik 1.0. Dengan data yang sama, PSO menyelesaikan permasalahan penjadwalan dengan 7 pelanggaran pada iterasi ke 50 dalam waktu 41.636 detik dengan fitness terbaik 0.111. nilai fitness GA menggunguli PSO, sebaliknya PSO memiliki standar deviasi yang cenderung lebih rendah dibandingkan PSO dengan artian hasil fitness yang dihasilkan PSO lebih stabil dibandingkan GA.

11. Algoritma klasifikasi c45 berbasis particle swarm optimization untuk prediksi hasil pemeilihan legislatif DPRD Karawang. (Nurrahman, 2017)

    Dalam penelitian ini, penulis akan memprediksi hasil pemilu legislatif dengan menggunakan predikto no urut parpol, suara sah partai, suara sah caleg dan jumlah perolehan kursi partai. Algoritma yang dipakai dalam penelitian tersebut yakni C45 yang dioptimasi dengan algoritma PSO. Hasil yang didapat yakni ketika menggunakan algoritma C45 menghasilkan akurasi sebesar 92.91% sedangkan ketika dikombinasikan dengan algoritma PSO menjadi 94.76%. Evaluasi menggunakan ROC curve yaitu berdasarkan nilai AUC, algoritma klasifikasi C45 bernilai 0.953 sedangkan C45 berbasis PSO bernilai 0.9561 dengan selisih 0.008 sehingga dapat disimpulkan algoritma PSO dapat meningkatkan akurasi C45.

2.2 Data Mining

Data mining merupakan sebuah istilah yang digunakan untuk menggambarkan sebuah kegiatan penemuan pengetahuan di dalam sebuah kumpulan data. Data mining merupakan proses yang menggunakan teknik statistik, matematikan, kecerdasan buatan dan pembelajaran berbasis mesin untuk mengekstrak dan mengidentifikasi informasi yang berguna beserta pengetahuan dari sebuah bank data yang besar.(Efraim Turban, Jay E. Aronson, 2007)

Semakin berkembangnya teknologi dan teknik dalam menganalisa, terdapat juga banyak pengertian dari data mining lainnya yang diungkapkan oleh beberapa ahli, diantaranya menurut Santoso, 2007 (Isidorus Cahyo Adi Prasetyo, 2016)menjelaskan bahwa data mining atau yang sering juga disebut knowledge discovery in database (KDD) merupakan kegiatan pengumpulan, pemakaian data historis untuk menemukan keteraturan, pola atau hubungan dalam set data berukuran besar. Keluaran datary proses data mining ini bisa digunakan untuk memperbaiki pengambilan keputusan dimasa depan. Kegiatan penggalian data ini dimaksudkan untuk menemukan pola yang menarik dari kumpulan data yang besar yang selanjutnya data tersebut disimpan dalam sebuah database, data warehouse atau penyimpanan informasi lainnya.

Istilah *data mining* dan *Knowledge Discovery in Database* (KDD) sering kali digunakan secara bergantian untuk menjelaskan proses penggalian informasi tersembunyi dalam suatu basis data yang besar. Sebenarnya kedua istilah tersebut memiliki konsep yang berbeda, tetapi berkaitan satu sama lain. Dan salah satu

tahapan dalam keseluruhan proses *Knowledge Discovery in* Database(KDD) adalah *data mining.*

Selain dari penjelasan diatas, terdapat penjelasan lain dari Suyanto (2017) menjelasakan bahwa Data mining merupakan gabungan dari sejumlah disiplin ilmu computer (ACM 2006), (Clifton, 2010) yang didefinisikan sebagai proses penemuan pola-pola baru dari kumpulan-kumpulan data sangat besar, meliputi metode-metode yang merupakan irisan dari artificial intelligence, machine learning, statistics dan database system (ACM 2006)

2.2.1 Kegunaan Data Mining

Secara umum kegunaan data mining dibagi menjadi dua, yaitu deskriptif dan prediktif. Data mining dikategorikan deskriftif jika digunakan untuk mencari pola-pola yang dapat dipahami manusia yang menjelaskan karakteristik data. Sedangkan prediktif berarti data mining digunakan untuk membentuk sebuah model pengetahuan yang digunakan untuk melakukan prediksi. (*Suyanto, 2017*)

Setiap aplikasi kelas data mining dibantu oleh pendekatan sebuah algoritma untuk mengekstrak suatu hubungan dalam sebuah data. Perbedaan kelas pendekatan algoritma tersebut dapat membantu dalam menyelesaikan masalah yang ditemui. Beberapa kelas tersebut adalah:

a. Classification

Mendefinisikan karakteristik dari sebuah grub atau membagi data ke beberapa jenis yang diketahui. Metode ini melibatkan sejumlah data yang diketahui kelas nya dan memetakan semua item tersebut kedalam sebuah set.

b. Clustering

Mengelompokkan data yang tidak diketahui label kelasnya kedalam sejumlah kelompok tertentu sesuai ukuran kemiripannya. Clustering dapat digunakan untuk mengidentifikasi kelas sesuai dengan kebutuhan.

c. Association

Mengidentifikasi hubungan antara kejadian yang terjadi dalam satu waktu. Pendekatan ini didasarkan pada analisis keranjang belanja. Pendekatan ini cocok menggunakan teknik statistik

d. Sequencing

Sama dengan pendekatan association, namun yang membedakan hanya di waktu terjadinya. Sequence mengukur waktu berdasarkan periode tertentu.

e. Regression

Menemukan suatu fungsi yang memodelkan data dengan kesalahana prediksi (galat) seminimal mungkin.

f. Forcasting

Mengestimasi nilai masa depan berdasarkan pola dalam sebuah data yang besar.

g. Anomaly detection

Mengidentifikasi data uang tidak umum, bisa berupa outlier (pencilan), perubahan atau deviasi yang mungkin sangat penting dan perlu investigasi lebih lanjut.

h. Association rule learning

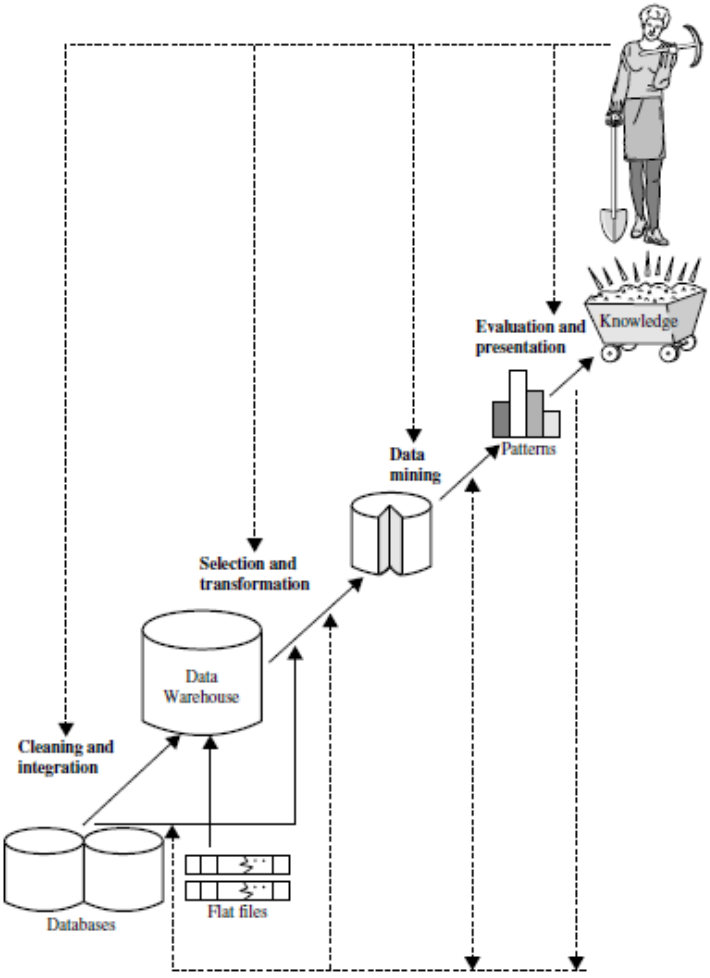
Bisa disebut juga pemodelan kebergantungan (dependency modelling), mencari relasi antar variable.

i. Summarization

Menyediakan representasi data yang lebih sederhana, meliputi visualisasi dan pembuatan laporan.

2.2.2 Tahapan Data Mining

Agar data yang ada dapat digunakan dan digali, data tersebut harus dipersiapkan terlebih dahulu agar dalam proses penggalian informasi tidak tejadi kesalahan. Terdapat beberapa teknik dalam mempersiapkan sebuah data yang akan digunakan dalam proses penggalian, diantara tahapan tersebut adalah:(Jiawei Han, Micheline Kamber and Jian Pei, 2006)



Gambar 2. Proses *mining data* untuk mendapatkan pengetahuan (Jiawei Han, Micheline Kamber and Jian Pei, 2006)

1. Cleaning dan integration

Data yang ada belum tentu bersih, sehingga perlu dibersihkan terlebih dahulu. Tujuan dari pembersihan ini adalah untuk menghilangkan noise dan data yang tidak konsisten. Setelah bersih baru data-data tersebut di integrasikan dengan satu penyimpanan sehingga tergabung menajadi satu data dan terpusat penyimpanannya.

2. Selection dan transformation

Pada tahap ini, data dan atribut yang akan digunakan diambil dari database untuk dianalisis. Selanjutnya data tersebut diubah menjadi bentuk yang tepat untuk selanjutnya dilakukan penggalian data untuk menemukan sebuah informasi.

3. Data mining

Tahap ini, data yang telah di seleksi selanjutnya dilakukan penggalian. Tujuannya adalah untuk menemukan pola atau informasi menarik dari data yang terpilih dengan menggunakan sebuah teknik atau metode. Banyak sekali teknik dalam memining data, tinggal di sesuaikan dengan tujuan dari me*mining* data tersebut.

4. Evaluation dan presentation

Pada tahap ini dilakukan identifikasi pola-pola yang benar-benar menarik dari hasil penggalian tadi. Setelah didapat polanya, perlu divisualisasikan dalam bentuk yang mudah dimengerti banyak orang.

2.3 Algoritma klasifikasi data mining

A. C4.5

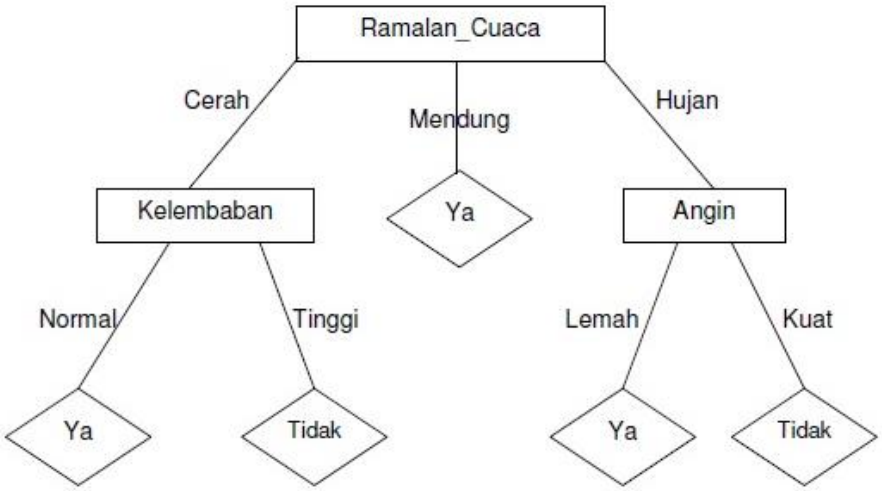
Algoritma C4.5 atau sering disebut juga *Decision Tree* merupakan metode klasifikasi dan prediksi yang sangat terkenal. Metode ini mengubah fakta yang sangat besar menjadi pohon keputusan yang representasikan aturan. Aturan dapat dengan mudah dipahami dengan bahasa alami. Selain iti juga dapat diekspresikan dalam bentuk bahasa basis data seperti SQL untuk mencari *record* padada kategori tertentu. Metode

ini juga berguna dalam mengeksplorasi data, menemukan hubungan tersembunyi antara sejumlah calon variabel input dengan sebuah variabel target. Karena *decision tree* memadukan antara eksplorasi data dan pemodelan. Metode ini digunakan untuk kasus dimana outputnya bernilai diskrit.(Gian Fiastantyo, 2009)

Proses pada *decision tree* adalah mengubah bentuk data (tabel) menjadi model pohon, mengubah model pohon menjadi *rule* dan menyederhanakannya. Sebuah model *decision tree* terdiri dari sekumpulan aturan untuk membagi sejumlah populasi yang heterofen menjadi lebih kecil, lebih homogen dengan memperhatikan pada variabel tujuannya. Variabel tujuan biasanya dikelompokkan dengan pasti dan lebih mengarah pada perhitungan probabilitas dari tiap-tiap record terhadap kategori tersebut atau untuk mengklasifikasiikan record dengan mengelompokkan dalam satu kelas.

Data dalam *decision tree* biasanya dinyatakan dalam bentuk tabel dengan atribut dan record. Atribut menyatakan suatu parameter yang dibuat sebagai kriteria dalam pembentukan pohon. Atribut ini juga memiliki nilai yang terkandung didalamnya yang disebut *instance*. Dalam *decision tree* setiap atribut akan menempati posisi simpul. Selanjutnya setiap simpul akan memiliki jawaban yang dibentuk dalam cabang-cabang, jawaban ini adalah instance dari atribut (simpul) yang dinyatakan. Pada saat penelusuran, pertanyaan pertama akan dinyatakan pada simpul akar. Selanjutnya akan dilakukan penelusuran ke cabang-cabang simpul akar dan simpul-simpul berikutnya. Penelusuran setiap simpul ke cabang-cabangnya akan berakhir ketika suatu cabang telah menemukan simpul kelas atau objek yang dicari.

Gambar 2.4Contoh pohon keputusan



Saat menyusun sebuah *decision tree* pertama yang harus dilakukan adalah menentukan atribut mana yang akan menjadi simpul akar dan atribut mana yang akan menjadi simpul selanjutnya. Pemilihan atribut yang baik adalah atribut yang memungkinkan untuk mendapatkan *decision tree* yang paling kecil ukurannya atau atribut yang bisa memisahkan objek menurut kelasnya. Secara heuristik atribut yang dipilih adalah atribut yang menghasilkan simpul yang paling bersih (purest). Ukuran purity dinyatakan dengan tingkat impurity dan menghitungnya dapat dilakukan dengan menggunakan konsep *entropy*. *Entropy* menyatakan impurity suatu kumpulan objek. Jika diberikan sekumpulan objek dengan label/output *y* yang terdiri dari objek berlabel 1, 2 sampa *n*, entropy *dari* objek dengan *n* kelas ini dapat dihitung dengan rumus berikut:

$$Entropy\ (S) = \sum_{i=l}^{n} -p_i * log2\ p_i \ldots\ldots\ldots\ (2.1)$$

keterangan:
S : himpunan kasus
A : atribut
N : jumlah partisi S
Pi : proporsi dari Si terhadap S

setelah itu, ada beberapa kriteria yang dibahas, yakni *information gain*, *gain ratio*, *indeks gini*.

1. *Information Gain*

Merupakan kriteria yang paling populer untuk pemilihan atribut. Information gain dapat dihitung dari output data atau variabel dependen *y* yang dikelompokkan berdasarkan atribut *A*, dinotasikan dengan gain (*y,A)*. Gain *(y,A)* dari atribut *A* relatif terhadap output data *y* adalah:

$$Gain\ (S,A) = Entropy\ (S\ ) - \sum_{i=l}^{n} \frac{|Si|}{|S|} * Entropy(Si) \ldots\ldots\ldots \quad (2.2)$$

Dimana nilai (A) adalah semua nilai yang mungkin dari atribut A dan y_c adalah subset dari *y* dimana *A* mempunyai nilai *c*.

2. *Gain Ratio*

Untuk menghitung gain ratio diperlukan suatu term split information. Split information dapat dihitung dengan formula sebagai berikut:

$$SplitInfo_A(D) = - \sum_{j=1}^{v} \frac{|Dj|}{|D|} x\ log2\ (\frac{Dj}{D}) \ldots\ldots\ldots \quad (2.3)$$

Selanjutnya gain ratio dihitung dengan cara:

$$GainRatio\ (A) = \frac{Gain(A)}{SplitInfo(A)} \ldots\ldots\ldots\ldots (2.4)$$

Adapun tahapan dalam membuat sebuah pohon keputusan dengan algoritma C4.5 adalah sebagai berikut:

1. Menyiapkan data training. Data training biasanya diambil dari data histori yang pernah terjadi sebelumnya dan sudah dikelompokkan ke dalam kelas-kelas tertentu.
2. Menentukan akar dari pohon. Akar diambil dari atribut yang terpilih dengan cara menghitung nilai gain pada masing-masing atribut dengan rumus persamaan 2.2 diatas. Nilai gain yang paling tinggi yang nantinya akan menjadi node pertama. Sebelum menghitung gain dari atribut, maka hitung hadulu nilai entropy nya dengan persamaan 2.1 diatas.
3. Ulangi langkah ke-2 hingga semua tupel terisi.
4. Proses partisi pohon keputusan akan berhenti saat:
    a. Semua tupel dalam node N mendapat kelas yang sama
    b. Tidak ada atribut didalam tupel yang dipartisi lagi.
    c. Tidak ada tupel didalam cabang yang kosong.

2.4 PSO (Particle Swarm Optimization)

Algoritma ini merupakan algoritma berbasis populasi yang mengeksploitasi individu dalam pencarian. Dalam PSO populasi disebut swarm dan individu disebut particle.

Tiap particle berpindah dengan kecepatan yang diadaptasi dari daerah pencarian dan menyimpan sebagai posisi terbaik yang pernah dicapai.

PSO didasarkan pada perilaku sosial sekawanan burung atau sekumpulan ikan. Algritma PSO meniru perilaku sosial organisme tersebut. Perilaku sosial terdiri dari tindakan individu dan pengaruh dari individu-individu lain dalam suatu kelompok. Kata partikel menunjukkan misalnya seekor burung dalam kawanan burung. Setiap individu atau partikel berprilaku secara terdistribusi dengan cara menggunakan kecerdasannya sendiri dan juga dipengaruhi perilaku kelompok kolektifnya. Dengan demikian jika suatu partikel atau seekor burung menemukan jalan yang tepat atau pendek menuju sumber makanan, sisa kelompok yang lain akan dapat segera mengikuti jalan tersebut meskipun lokasi mereka jauh di kelompok tersebut.

Metode optimasi yang didasarkan pada *swarm intelligence* ini disebut algoritma *behaviorully inspired* sebagai alternatif dari algoritma genetika yang sering disebut *evolution-based procedures.* Algoritma PSO ini awalnya diusulkan oleh *J.Kennedy* dan *R.C.Eberhart*(Eberhart & Kennedy, 2002). Dalam konteks optimasi multivariabel, kawanan diasumsikan mempunyai ukuran tertentu atau tetap dengan setiap partikel posisi awalnya terletak di suatu lokasi yang acak dalam ruang multidimensi. Setiap partikel bergerak dalam ruang tertentu dan mengingat posisi terbaik yang pernah dilalui atau ditemukan terhadap sumber makanan atau nilai objektif. Setiap partikel menyampaikan informasi atau posisi bagusnya kepada partikel uan lainnya dan menyesuaikan posisi dan kecepatan masing-masing berdasarkan informasi yang diterima mengenai posisi yang bagus tersebut. Sebagai

contoh, perilaku burung-burung dalam kawanan burung. Meskipun setiap burung mempunyai keterbatasan dalam hal kecerdasan, biasanya ia akan mengikuti kebiasaan sebagai berikut:

1. Seekor burung tidak berada terlalu dekat dengan burung lain
2. Burung tersebut akan mengarahkan terbangnya ke arah rata-rata keseluruhan burung
3. Akan memposisikan diri dengan rata-rata posisi burung yang lain dengan menjaga, sehingga jarak antar burung dalam kawanan tidak terlalu jauh.

Dengan demikian perilaku kawanan burung akan didasarkan pada kombinasi dari 3 faktor berikut:

1. Kohesi - terbang bersama
2. Separasi – terbang jangan terlalu dekat
3. Penyesuaian (*alignment*) – mengikuti arah bersama

Jadi algoritma PSO dikembangkan dengan berdasar pada model berikut:

1. Ketika seekor burung mendeteksi target atau makanan (atau bisa minimum atau maksimum suatu fungsi tujuan) secara cepat mengirim informasi kepada burung-burung yang lainnya dalam kawanan.
2. Burung yang lainnya akan mengikuti arah menuju ke makanan tetapi tidak secara langsung
3. Ada komponen yang tergantung pada pikiran burung, yaitu memori tentang apa yang sudah dilewati pada waktu sebelumnya.

Keuntungan dari algoritma PSO ini adalah sebagai berikut:

1. PSO berdasar pada kecerdasan (*intelligence*). Ini dapat diterapkan ke dalam kedua penggunaan dalam bidang teknik dan riset ilmiah.
2. PSO tidak punya *overlap* dan kalkulasi mutasi. Pencarian dapat dilakukan oleh kecepatan dari pertikel. Selama pengembangan beberapa generasi, kebanyakan hanya partikel yang optimis yang dapat mengirim informasi ke partikel yang lain dan kecepatan dari pencarian adalah sangat cepat.
3. Perhitungan didalam algoritma PSO sangat sederhana, menggunakan kemampuan optimisme yang lebih besar dan dapat diselesaikan dengan mudah.
4. PSO memakai kode/jumlah yang riil dan itu diputuskan langsung dengan solusi dan jumlah dimensi tetap sama dengan solusi yang ada.

Beberapa kerugian dari algoritma PSO ini adalah:

1. Metode mudah mendapatkan optimal parsial (sebagian) yang mana menyebabkan semakin sedikit ketepatan untuk peraturan tentang arah dan kecepatan.
2. Metode tidak bisa berkembang dari permasalahan sistem yang tidak terkoordinir seperti solusi dalam bidang energi dan peraturan yang tidak menentu dibidang energi.

Model ini akan disimulasikan dalam ruang dengan dimensi tertentu dengan sejumlah iterasi sehingga di setiap iterasi, posisi partikel akan semakin mengarah ke target yang dituju (minimum atau maksimum fungsi). Hal ini dilakukan hingga maksimum iterasi dicapai atau bisa digunakan kriteria penghentian yang lain.

Setiap partikel dalam PSO juga dikaitkan dengan kecepatan partikel terbang melalui ruang pencarian dengan kecepatan yang dinamis disesuaikan untuk perilaku historis mereka. Oleh karena itu, partikel memiliki kecenderungan untuk terbang menuju daerah pencarian yang lebih baik selama proses pencarian.

Didalam penelitian yang dilakukan oleh (Amperiana, 2015)dijelaskan juga bahwa algoritma PSO dalam pencarian solusi dilakukan oleh suatu populasi yang terdiri dari beberapa partikel. Populasi dibangkitkan secara acak (*random*) dengan batasan nilai terkecil (X_{min}) dan nilai terbesar (X_{max}). Setiap partikel merepresentasikan posisi atau solusi dari permasalahan yang dihadapi. Setiap partikel melakukan pencarian solusi yang optimal dengan melintasi ruang pencarian (*search space*). Hal ini dilakukan dengan cara setiap partikel melakukan peneyesuaian terhadap posisi partikel terbaik dari partikel tersebut (*local best*) dan penyesuaian terhadap posisi partikel terbaik dari seluruh kawanan (*global best*) selama melintasi ruang pencarian. Jadi penyebaran pengalaman atau informasi terjadi di dalam partikel itu sendiri dan antara suatu partikel dengan partikel terbaik dari seluruh kawanan selama proses pencarian solusi. Setelah itu, dilakukan proses pencarian untuk mencari posisi terbaik setiap partiukel dalam sejumlah iterasi tertentu sampai didapatkan posisi yang relatif tetap (*steady*). Dengan kata lain, nilai fungsi mulai konvergen atau iterasi telah mencapai batas yang telah ditetapkan. Pada setiap iterasi, setiap solusi yang direpresentasikan oleh posisi partikel dievaluasi performanya dengan cara memasukkan solusi tersebut kedalam fungsi kesesuaian (*fitness function*).

Penelitian (Kusmarna, Wardhani, & Safrizal, 2015) juga menjelaskan tentang algoritma PSO yang terdiri dari tiga tahap, yaitu pembangkitan posisi serta kecepatan partikel, *update velocity* (pembaruan kecepatan), *update position* (pembaruan posisi.). Pertama posisi X_x^i dan kecepatan V_x^i dari kumpulan partikel dibangkitkan secara *random*menggunakan batas atas (X_{max}) dan batas bawah (X_{min}) dari *design variabel* seperti yang ditunjukkan pada persamaan dibawah ini:

$$X_x^i = X_{max} + rand\ (X_{max} - X_{mix}) \quad \ldots\ldots (2.5)$$

$$V_x^i = X_{max} + rand\ (X_{max} - X_{mix}) \quad \ldots\ldots (2.6)$$

Dimana:

X_x^i = posisi awal

V_x^i = kecepatan awal

X_{max} = batas bawah

X_{mix} = batas atas

Rand = nilai acak antara 0 dan 1

Adapun persamaan matematis yang menggambarkan untuk menghitung pembaruan stasus perpindahan posisi dan kecepatan partikel (*update velocity dan* posistion) adalah sebagai berikut:

$$V_i(t) = V_i\ (t\text{-}1) + c_1 r_1 [X_{pbest\ i} - X_i\ (t)] + c_2 r_2 [X_{Gbest} - X_i\ (t)] \quad \ldots\ldots (2.7)$$

$$X_i\ (t) = X_i\ (t\text{-}1) + V_i\ (t) \quad \ldots\ldots (2.8)$$

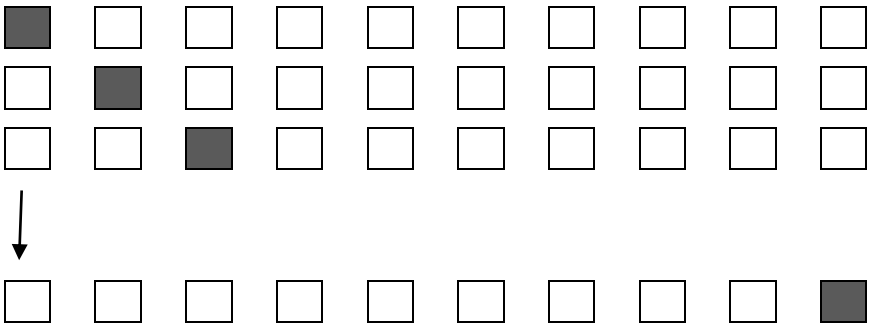
Dimana:

$V_i(t)$	:kecepatan partikel *i* saat iterasi *t*
$X_i(t)$	: posisi partikel *i* saat iterasi *t*
c_1 *dan* c_2	: *learning rate* untuk kemampuan individu (cognitive) dan pengaruh sosial
r_1 *dan* r_2	: bilangan random yang berdistribudi uniformal dalam interval 0 dan 1
$X_{pbest\ i}$	: posisi terbaik partikel *i*
X_{Gbest}	: posisi terbaik global

Persamaan (2.7) digunakan untuk menghitung kecepatan partikel yang baru berdasarkan kecepatan sebelumnya, jarak antara posisi saat ini dengan posisi terbaik partikel (*Pbest*) dan jarak antara posisi saat ini dengan posisi terbaik kawanan (*Gbest*). Kemudian partikel terbang menuju posisi yang baru berdasarkan persamaan (2.8).

2.5 Cross Validation

Criss Validation adalah sebuah tekniku untuk mengambil sampel dari sebuah data yang banyak secara acak yang menjamin setiap kemunculan data yang diamati sama dengan jumlah data latih dan kemunculan pada data *testing* hanya sekali. Dalam teknik validasi ini, kita menetapkan jumlah partisi atau *fold.* Standar yang biasa digunakan untuk memperoleh estimasi kesalahan terbaik adalah 10 kali partisi *(tenfold cross-validation)*. Data dibagi secara acak menjadi 10 bagian dengan perbandingan yang sama kemudian *error rate* dihitung bagian demi bagian, selanjutnnya *error rate* secara keseluruhan diperoleh dari menghitung rata-rata *error rate* dari 10 bagian tersebut. Ilustrasi dari cara kerja taknik ini dapat dilihat pada gambar dibawah ini:



Data latih

Data *testing*

Gambar 2.5 Ilustrasi *tenfold cross-validation*

2.6 Confusion Matrix

Untuk melakukan evaluasi terhadap model klasifikasi berdasarkan perhintungan objek testing mana yang diprediksi benar dan tidak benar. Perhitungan ini di tabulasikan kedalam tabel yang disebut *confusion matrix*

Confusion matrix merupakan *dataset* yang memiliki 2 kelas, kelas yang satu sebagai positif dan kelas yang lainnya sebagai negatif. Terdiri dari 4 sel yaitu *true positive (TP), False Positif (FP), True Negative (TN)* dan *False Negative (FN).*

CLASSIFICATION	**Prediction Class**		
OBSERVED CLASS		**Class = *YES***	***Class = NO***
	Class = *YES*	a (true positive – TP)	b (false negative – FN)
	Class = NO	c (false positive – FP)	d (true negative – TN)

Gambar 2.1 Confusion matrix untuk 2 model kelas

Untuk menghitung akurasi menggunakan rumus sebagai berikut:

$$Accuracy = \frac{a+d}{a+c+d+b} \text{ atau } \frac{TP+TN}{TP+FP+TN+FN}$$

Model C4.5	Kelas yang diprediksi	
Kelas yang diamati	250	45
	5	200

Gambar 2.2 Contoh Confusion matrix

Dari tabel diatas dapat dilakukan pengukuran akurasi model C4.5 sebagai berikut:

2.7 Kurva ROC (*Receiver Operation Characteristic*)

Kurva ROC menunjukkan visualisasi dari akurasi model dan perbandingan perbedaan antar model klasifikasi. Kurva ROC mengekspresikan *confusion matrix.* ROC adalah grafik dua dimensi dengan *false positive* sebagai garis horiozontal dan *true positive* untuk mengukur perbedaan performasi metode yang digunakan(Gorunescu, 2011). Kurva ROC merupakan teknik untuk mevisualisasikan dan menguji kinerja pengklasifikasian. Model klasfikasi yang lebih baik yang menunjukkan ROC lebih besar.

Selain ROC, terdapat pula kurv AUC (*area under* cureve) yaitu kurva yang berada di bawah area kurva ROC. Semakin tinggi AUC maka semakin baik. Model yang baik mempunyai AUC didekat angka 1 sedangkan yang mendekati angka 0 maka modelnya tidak baik. Performa keakurasian AUC dapat diklasifikasikan menjadi lima kelompok, yaitu:

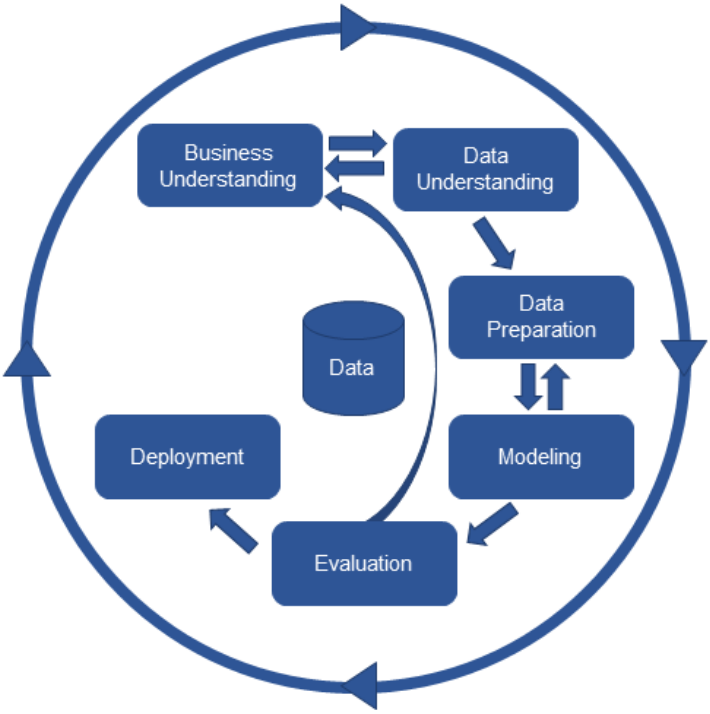
a. 0.90 – 1.00 = unggul
b. 0.80 – 0.90 = baik
c. 0.70 – 0.80 = cukup
d. 0.60 – 0.70 = kurang
e. 0.50 – 0.60 = gagal

2.8 Metodologi CRISP-DM

*Cross Industry Standard Process for Data Mining*atau yang sering disebut dengan CRISP-DM adalah standarisasi *data mining* yang disusun oleh 4 penggagas *data mining market*, yaitu *Daimler Chrysler (Daimler-Benz), SPSS (ISL),NCR and OHRA*, lalu dikembangkan pada berbagai workshops antara tahun 1997-1999. Terdapat lebih dari 300 organisasi berkontribusi dalam proses modelling metode in dan akhirnya CRISP-DM 1.0 dipublikasikan pada tahun 1999. (Watson et al., 2000)

CRISP-DM merupakan sebuah metodology dalam data mining yang bersifat menyeluruh dan menyediakan model dalam suatu proses. Metodologi ini dapat digunakan dari level data mining pemula hingga level expert dengan menyediakan *blueprint* yang lengkap untuk suatu proyek data mining. CRISP-DM membuat membagi siklus dalamdata mining menjadi 6 tahap yang mana tahapan tersebut akan dijelaskan di bawah ini:

Gambar 2.3 Siklus CRISM-DM



Business understanding

atau pemahaman bisnis atau bisa juga disebut sebagai pemahaman organisasi karena dari segi organisasi pun harus diketahui secara menyeluruh bukan hanya berfokus kepada bisnis. Langkah ini penting untuk dilakukan jika akan melakuka penambangan data, karena melalui fase ini, segala kebutuhan informasi tambahan akan bisa didapatkan dari kumpulan data yang akan ditambang(North, 2012). Pada fase ini dibutuhkan pemahaman tentang substansi dari kegiatan data mining yang akan dilakukan, kebutuhan dari perspektif organisasi maupun bisnis. Kegiatan ini antara lain menentukan sarana dan tujuan bisnis, memahami situasi bisnis, menetukan tujuan data mining dan membuat perencanaan strategi serta jadwal penelitia.

Data Understanding

atau pemahaman data adalah fase mengumpulkan data awal, mempelajari data untuk bisa mengenal data yang akan dipakai. Fase ini mencoba mengidentifikasikan masalah yang berkaitan dengan kualitas data, mendeteksi subset yang menarik dari data untuk dibuat hipotesa.

Fase ini termasuk dalam masa persiapan dan banyak yang melewatkanya. Pada zaman dahulu ketika perangkat komputer belum ditemukan, segala aktivitas dalam pengolahan data masih berbasis kertas dan disimpan didalam lemari arsip yang telah dipersiapkan. Sehingga sifat data pada zaman dulu bersifat sentralisasi, karena ketika seseorang membutuhkan sebuah data , maka orang tersebut tinggal datang ke kantor mereka dan menghubungi bagian administrasi umum dan meminta data yang dibutuhkan dan hal tersebut tinggal dicari di lemari arsip yang ada.

Berbeda dengan zaman sekarang yang segala sesuatu sudah menggunakan perangkat komputer dan banyaknya media penyimpanan yang tersebar. Hal ini menyebabkan data tidak lagi bersifat sentral atau terpusat. Data menjadi tersebar dimana-dimana dan diberbagai media penyimpanan yang ada di komputer maupun yang tidak.

Hal ini menjadi permasalahan baru yang dihadapi dalam sebuah organisasi, ketika akan mencari informasi baru dari data yang dimiliki, namun keberadaan data masih tersebar di berbagai unit bidang dan bisa jadi di masing-masing unit bidang tersebut justru tidak mengetahui tentang data apa yang mereka miliki.

Namun kegiatan pengumpulan saja belum cukup. Terkait dengan asal muasal data dan siapa yang mengumpulkan serta dengan metode apa dalam pengumpulanya pun harus menjadi konsen oleh orang yang mengumpulkan data. Terkadang juga harus

bertemu dengan para ahli di bagian unit bisnis tersebut untuk memverifikasi tentang seluk beluk data tersebut serta memverifikasi tentang keakuratan dan keandalan data yang ada. Data yang tidak lengkap dan akurat dapat berpengaruh terhadap proses penambangan yang mana akan memunculkan informasi yang tidak akurat juga serta mempengaruhi keputusan yang harus diambil sehingga menghasilkan keputusan yang salah.

Data Preparation

atau persiapan data. Fase ini sering disebut sebagai fase yang padat karya. Aktivitas yang dilakukan diantaranya memilih *table* dan *field* yang akan ditransformasikan ke dalam database baru untuk bahan data mining (set data mentah). Bisa juga menggabungkan dua atau lebih data yang didapat dari berbagai sumber, mengurangi variabel-variabel yang tidak penting keberadaannya dan hanya mengambil variabel atau *field*yang berguna dalam proses penambangan.

Modeling

adalah fase menentukan teknik data mining yang digunakan, menentukan *tools data mining*, teknik, algoritma serta menentukan parameter dengan nilai yang optimal.

Terdapat 2 tipe dasar model dalam data mining, yakni *classify* and *predict*.

Seperti contoh model *decision tree. Decision tree* merupakan model prediksi yang digunakan untuk menentukan atribut yang terkuat yang akan memberikan pengaruh terhadap hasil dari dataset yang diberikan. Jadi model *decision tree* bersifat prediktif

namun dapat juga membantu kita untuk mengklasifikasikan terhadap data yang kita miliki.

Model membantu kita untuk mengklasifikasikan dan memprediksi berdasarkan pola model yang terdapat didata yang kita miliki. Model biasanya hanya berbentuk sederhana, hanya berisi satu proses atau terdiri dari sub-sub proses. Terlepas dari bentuk mereka, model merupakan tempat dimana data bergerak dari tahap persiapan dan interprestasi.

Evaluation

adalah fase interprestasi terhadap hasil data mining yang ditunjukkan dalam proses pemodelan pada fase sebelumnya. Evaluasi dilakukan secara mendalam dengan tujuan menyesuaikan model yang didapat agar sesuai dengan sasaran yang ingin dicapai dalam fase pertama.

Dalam proses analisis data, kemungkinan muncul kesalahan pasti ada.. seperti tidak menemukan pola yang menarik dari data yang akan ditambang. Bisa disebabkan karena model tidak diatur dengan baik pada saat pembangunan model atau bisa juga menggunakan teknik yang salah atau mungkin tidak ada sesuatu yang menarik dari data yang diberikan.

Evaluasi dapat dilakukan dengan menggunakan sejumlah teknik, baik matematis maupun logic. Selain itu juga evaluasi model juga harus memasukkan aspek manusia. Ketika *user* mendapatkan pengalaman dan keahlian dibidangnya, mereka akan

memiliki pengetahuan operasional yang mungkin tidak dapat diukut secara matematis.

Deployment

atau penyebaran adalah fase penyusunan laporan atau presentasi dari pengetahuan yang didapat dari evaluasi pada proses data mining. Aktivitasi ini juga termasuk mengatur model secara otomatis, melakukan pertemuan dengan pelanggan dari model yang direkomendasikan, mengintegrasikan dengan manjemen atau operasional dari sistem informasi yang ada, menemukan teknik pembelajaran baru dari model yang ada untuk mengembangkan tingkat akurasi dan kemampuan serta memantau dan mengukur hasil model yang telah diimplementasikan.

2.9 RapidMiner

RapidMiner merupakan platform perangkat lunak yang dikembangkan untuk keperluan persiapan data, mesin pembelajaran, penambangan teks dan analisis prediktif (RapidMiner, 2012).

RapidMiner sebelumnya dikenal sebagai YALE (*Yet Another Learning Enviorenment*) yang mulai dikembangkan pada tahun 2001 oleh *Raif Klinkenberg*, *Ingo Mierswa* dan *Sion Fischer* dari unit kecerdasan buatan universitas teknik dortmund yang ditulis menggunakan bahasa pemrograman java. Mulai tahun 2006, pengembangan selanjutnya didorong oleh Rapid-I sebuah perusahaan yang didirikan oleh *Ingo Mierswa* dan *Raif Klinkenberg* pada tahun yang sama. Pada tahun 2007, nama perangkat lunak yang sebelumnya bernama YALE berubah menjadi

RapidMiner. Pada tahun 2013 perusahaan Rapid-I melakukan *rebranding* dari yang semula bernama Rapid-I menjadi RapidMiner.

RapidMiner menyediakan prosedur data mining dan *machine eraning* termasuk ETL (*Extraction, Transformation, Loading*), data *prepocessing* dan evaluasi. Proses *data mining* tersusun atas operator-operator yang *netstable,* dideskripsikan dengan XML dan dibuat dengan GUI.

Dalam RapidMiner akan sering kita mendengar istilah atribut, atribut target. Atribut merupakan karakteristik atau fitur dari data yang menggambarkan sebuah proses atau situasi, contoh id. Sedangkan atribut target merupakan atribut yang menjadi tujuan untuk diisi oleh proses data mining.

2.10Kerangka pemikiran

Dalam penelitian ini, permasalahan yang diangkat yakni peningkatan tingkat keakuratan algoritma C4.5 dalam memprediksi kelulusan siswa peserta seleksi bersama masuk perguruan tinggi negeri. Untuk meningkatkan akurasi tersebut, dipilihlah algoritma *particle swarm optimization* (PSO) dalam seleksi atribut. Desain penelitian ini menggunakan metodologi CRISP-DM sedangkan aplikasi untuk membantu dalam pengolahan data menggunakan RapidMiner. Pengujian terhadap hasil kinerja algoritma C4.5 dan C4.5 berbasis PSO menggunakan metode *Cross Validation,* tingkat akurasi algoritma diukur dengan *Confusion Matrix* dan AUC dengan kurva ROC. Dari hasil perbandingan nilai akurasi, maka akan diketahui dampak penerapan PSO dalam algoritma C4.5

Gambar 2.4 Kerangka pemikiran penelitian