BAB 2
TINJAUAN PUSTAKA DAN DASAR TEORI

2.1 KAJIAN PUSTAKA

Putra, I Gede Jana Adi, dkk (2019) mengidentifikasi Gaya Belajar Peserta Didik menggunakan *Bayesian Network*, klasifikasi data dilakukan wawancara satu persatu pada peserta didik kelas XI MIA 1 SMAN 1 Kuta dengan jumlah peserta didik 34 orang. Peserta didik diberikan kesempatan mempelajari materi, lalu diberikan evaluasi *post test* maksimal 3 kali pada materi yang telah dipelajari untuk mencapai nilai minimal 74. Tiap pertemuan disediakan 3 jenis konten pembelajaran yaitu teks, audio dan video. Deteksi gaya belajar dilakukan setelah pertemuan 3 dari 5 pertemuan yang dilakukan sebagai bentuk uji coba. Rata-rata nilai hasil belajar siswa dengan adaptive learning 82.00, nilai tertinggi 100, dan nilai terendah 47.00. Hasil identifikasi penelitian terdapat 14 peserta didik dengan gaya belajar tekstual, 1 audio, dan 1 visual [4].

Sasongko, Theopilus Bayu dan Oki Arifin (2019) membandingkan performansi algoritma SVM dan *Naïve bayes kernel density* menggunakan *forward selection* serta tanpa *forward selection*. Pengumpulan data peminatan dilakukan di dua sekolah berbeda, data pertama dinamakan *dataset* ABC memiliki 11 atribut meliputi nama, nilai Matematika, IPA, MIPA, IPS, tes psikologi IQ, logika rasional, konkrit operasional, abstrak konseptual, analisa *sintesa*, logika verbal, logika *numeric*, dan daya ingat, jumlah *records* 280 siswa. Siswa IPA 150 orang dan siswa IPS 138 orang. Proses *training* algoritma SVM didahului dengan normalisasi menggunakan *min-max normalization* dengan rentang +1 dan -1. Proses *training* algoritma *Naïve bayes classifier kernel density* tidak menggunakan normalisasi karena *Naïve bayes kernel density* dapat menangani nilai atau data diskrit dengan baik. Proses training algortima *Naïve bayes classifier kernel density* menggunakan *estimation* mode *greedy number of kernel* 0.0-100.0 rentang 10.0. 10 *k-fold cross validation* membagi menjadi *data training* dan *data testing*. Model nilai akurasi paling tinggi digunakan untuk melakukan klasifikasi *dataset* peminatan sekolah kedua yang dinamai *dataset* XYZ jumlah *records* 288, 11 variabel diantaranya yaitu

nilai Matematika, IPA, tes akademik, IQ, logika rasional, konkrit, operasional, abstrak konseptual, nalisa sintesa, logika verbal, logika *numeric*, dan daya ingat. Pengujian model klasifikasi dataset XYZ menggunakan metode akurasi dan nilai AUC. *Dataset* peminatan ABC menghasilkan nilai akurasi terbesar oleh *kernel anova* 98.21% menggunakan algoritma SVM. Nilai akurasi terbesar oleh *kernel anova* 99.29% menggunakan algoritma FS-SVM. Metode *forward selection* dapat meningkatkan nilai performansi pada semua kernel yang diuji, terutama pada NBC nilai akurasi 98.21%. Nilai akurasi NBC menurun ketika *number of kernel* naik di angka 10.0. Hasil akurasi tertinggi *dataset* ABC algoritma FS-SVM berbasis *kernel anova* parameter C 10.0 sebesar 99.29%. Sedangkan *dataset* XYZ algoritma FS-SVM berbasis kernel anova parameter C 10.0 nilai akurasi 95.17% dan nilai AUC 0.956 [5].

Arifin, Oki dan Theopilus Bayu Sasongko (2018) membandingkan tingkat performansi SVM dan *Naïve Bayes Classifier* pada klasifikasi jalur minat SMA menggunakan data dua sekolah SMA sebanyak 288 siswa diambil pada tahun ajaran 2013-2014. *Dataset* pengujian pertama dinamakan dataset penjurusan ABC terdiri dari data nilai psikologi (IQ, logika rasional, konkrit operasional, konkrit operasional, abstrak konseptual, analisa *sintesa*, logika verbal, logika *numeric*, daya ingat), nilai UN jenjang sebelumnya (Matematika, IPA), dan nilai rata-rata raport kelas X (IPA, IPS). Siswa berlabel jurusan IPA berjumlah 150 siswa dan siswa berlabel jurusan IPS berjumlah 138 siswa. *Dataset* pengujian kedua dinamakan dataset penjurusan XYZ berjumlah 280 siswa, terdiri dari data nilai psikologi (IQ, logika rasional, konkrit operasional, abstrak konseptual, analisa sintesa, logika verbal, logika numeric, daya ingat), nilai UN jenjang sebelumnya (Matematika, IPA), nilai raport (Matematika, IPA), dan nilai tes akademik. Tingkat performansi yang diukur dengan akurasi, presisi, *recall*, dan nilai *Area Under Curve* (AUC) algoritma SVM menggunakan *kernel anova* dan faktor pinalti (C) sebesar 5.0 lebih besar daripada performansi algoritma *Naïve Bayes Classifier* [6].

Purnanditya, Bondhan Arya dan Ahmad Zainul Fanani (2015) menerapkan *Forward Selection* menggunakan algoritma *Naïve Bayes* untuk menentukan *attribute* yang berpengaruh pada klasifikasi kelulusan, data yang digunakan data kualitatif. Data dalam penelitian berasal dari IAsol UNAKI Fakultas Ilmu

Komputer tahun ajaran 2008-2011. Atribut berupa Nomor Induk Mahasiswa (NIM), nama, jurusan, umur, jenis kelamin, daerah asal, status pernikahan, status pekerjaan, kelompok atau jenis beasiswa, indeks prestasi dari semester 1-9, IPK, jumlah sks yang ditempuh, dan jenis konsentrasi jalur peminatan. *Dataset* memiliki 3 *class* atau 3 kategori kelulusan, dengan data yang memiliki 240 *record* dan 21 *attribute*. Pengujian menggunakan data sampel diambil dari IAsol *dataset* dengan 2 *label class* (tepat dan terlambat), 10 *record* (7 *class* tepat dan 3 *class* terlambat) dan 21 *attribute*. Metode *forward selection* dapat mereduksi dimensi *dataset* yang besar dan membantu meningkatkan hasil akurasi klasifikasi *Naïve bayes*. Penggunaan metode *forward selection* pada algoritma *Naïve bayes* lebih akurat dan efektif dalam mengklasifikasikan kelulusan mahasiswa dari dataset yang bersifat *class imbalance* menghasilkan tingkat akurasi sebesar 99.17%, memperoleh *attribute* yang berpengaruh yaitu kelompok, IP Semester 1, IP Semester 3, IP Semester 9, dan jika tidak menggunakan metode *forward selection* hanya menghasilkan nilai akurasi 95.83% [7].

K. Wabang, O. D. Nurhayati and Farikhin (2023) menerapkan algoritma *naïve bayes* untuk mengklasifikasi pengaduan masyarakat. Data yang digunakan pada penelitian terbagi menjadi dua yaitu data *training* dan data *testing* dengan perbandingan 80% data latih yaitu 362 *record* dan 20% data uji yaitu 91 *record*. Pengaduan atau laporan masyarakat terbagi menjadi tiga kelas, yaitu laporan sederhana, laporan sedang, dan laporan berat. Data dalam penelitian meliputi 5 atribut sebagai variabel bebas (x) terdiri dari jumlah masalah, jumlah instansi yang dilaporkan, lokasi yang dilaporkan, penerima manfaat, dan isu/perhatian publik. Label/kelas yang digunakan adalah klasifikasi laporan sebagai variabel terikat (y). Proses klasifikasi dilakukan dengan menggunakan penilaian bobot dari setiap pengaduan/laporan dengan menggunakan 5 (lima) atribut. Hasil penelitian dengan menerapkan algoritma *naïve bayes classifier* memberikan nilai akurasi yang tinggi sebesar 92%. Selain itu, rata-rata nilai *precision, recall*, dan *f1-score* masing-masing adalah 91%, 93%, dan 92% [8].

2.2 DASAR TEORI

2.2.1 Psikologi

Secara etimologi kata *psikologi* berasal dari bahasa Yunani Kuno *psyche* dan *logos*. Kata *psyche* berarti "jiwa, roh, atau sukma", sedangkan kata *logos* berarti "ilmu". Jadi, *psikologi*, secara harfiah berarti "ilmu jiwa", atau ilmu yang objek kajiannya adalah jiwa. Ketika psikologi masih berada atau merupakan bagian dari ilmu filsafat, definisi psikologi adalah ilmu yang mengkaji jiwa masih dipertahankan. Namun, istilah ilmu jiwa tidak digunakan lagi karena bidang ilmu ini memang tidak meneliti jiwa atau roh atau sukma, sehingga istilah itu kurang tepat. Dalam perkembangan lebih lanjut, psikologi lebih membahas atau mengkaji sisi-sisi manusia dari segi yang bisa diamati. Dalam hal ini "jiwa" atau "keadaan jiwa" hanya bisa diamati melalui gejala-gejalanya seperti orang yang sedih akan berlaku murung, dan orang yang gembira tampak dari gerak-geriknya yang riang atau dari wajahnya yang berbinar-binar. Tidak jarang kita jumpai seseorang yang sebenarnya sedih tetapi tetap tersenyum. Atau seseorang yang sebenarnya jengkel atau marah tetap tenang atau malah tertawa. Psikologi lazim diartikan sebagai satu bidang ilmu yang mencoba mempelajari perilaku manusia. Caranya adalah dengan mengkaji hakikat rangsangan, hakikat reaksi terhadap rangsangan itu, dan mengkaji hakikat proses-proses akal yang berlaku sebelum reaksi itu terjadi. Para ahli psokologi cenderung untuk menganggap psikologi sebagai suatu ilmu yang mencoba mengkaji proses "akal manusia" dan segala manifestasinya yang mengatur perilaku manusia itu. Tujuan pengkajian akal ini adalah untuk menjelaskan, memprediksikan, dan mengontrol perilaku manusia [9].

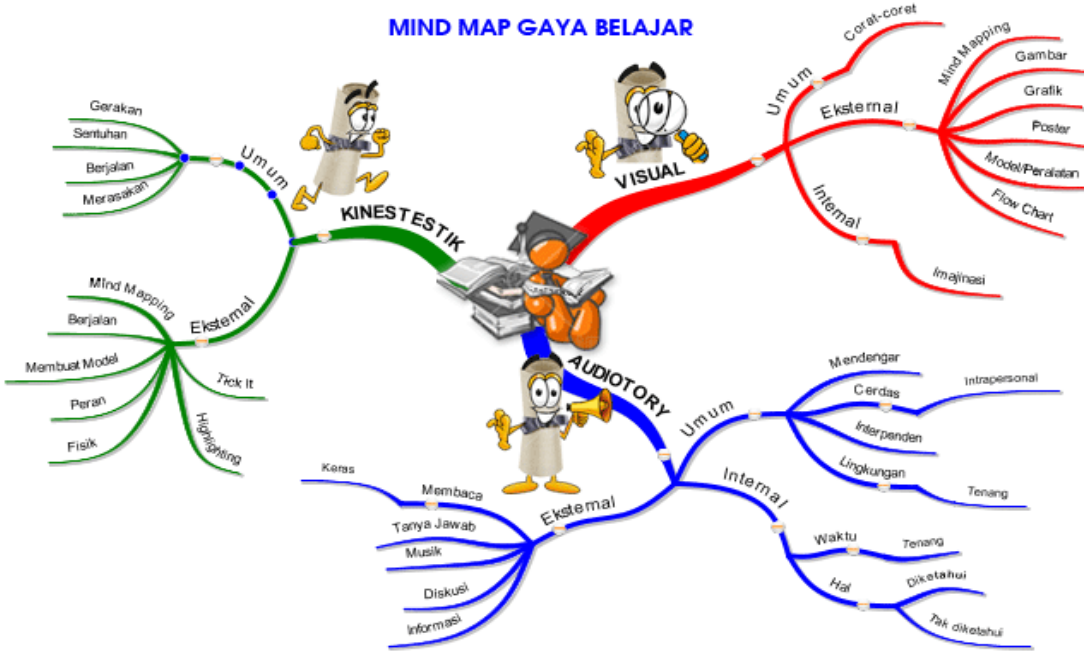
2.2.2 Tes Psikologi

Tes psikologi adalah sarana yang penting untuk melakukan penelitian secara psikologis. Terdapat banyak jenis tes psikologis, contohnya adalah tes dimana responden hanya perlu menjawab iya atau tidak. Bahkan tes lain dirancang sedemikian rupa sehingga peserta tes harus merespon dalam *virtual reality environment*. Beberapa tes psikologi berbasis computer, tetapi terdapat juga tes psikologi yang membutuhkan bertahun-tahun latihan dan pengalaman. Terlepas dari perbedaan di atas, semua tes psikologi dianggap memiliki satu kesamaan yaitu

sarana tersebut digunakan oleh psikolog untuk mengumpulkan data tentang manusia [10].

2.2.3 Gaya Belajar

De Poter & Hernacki (1999) menjelaskan secara umum gaya belajar manusia dibedakan ke dalam tiga kelompok besar gaya belajar, yaitu gaya belajar visual, gaya belajar auditorial, dan gaya belajar kinestetik [3].



Gambar 2. 1 Gaya Belajar *[11]*

2.2.3.1 Gaya Belajar Visual

Gaya belajar visual adalah gaya belajar dengan cara melihat, mengamati, memandang, dan sejenisnya. Kekuatan gaya belajar visual terletak pada indera penglihatan. Bagi individu yang memiliki gaya belajar ini, mata adalah alat yang paling peka untuk menangkap setiap gejala atau stimulus (rangsangan) belajar. Ciri-ciri individu yang memiliki tipe gaya belajar visual yaitu [3]:

1. Menyukai kerapian dan ketrampilan.
2. Jika berbicara cenderung lebih cepat.
3. Suka membuat perencanaan yang matang untuk jangka panjang.
4. Sangat teliti sampai ke hal-hal yang detail sifatnya.
5. Mementingkan penampilan baik dalam berpakaian maupun presentasi.

6. Lebih mudah mengingat apa yang didengar.
7. Mengingat sesuatu dengan penggambaran (asosiasi) visual.
8. Tidak mudah terganggu dengan keributan saat belajar.
9. Pembaca yang cepat dan tekun.
10. Lebih suka membaca sendiri dari pada dibacakan orang lain.
11. Tidak mudah yakin atau percaya terhadap setiap masalah sebelum secara mental merasa pasti.
12. Suka mencorat-coret tanpa arti selama berbicara di telepon atau dalam rapat.
13. Lebih suka melakukan pertunjukan (demonstrasi) daripada berpidato.
14. Lebih menyukai seni daripada musik.
15. Seringkali mengetahui apa yang harus dikatakan akan tetapi tidak pandai memilih kata-kata.
16. Kadang-kadang suka kehilangan konsentrasi ketika mereka ingin memperhatikan.

2.2.3.2 Gaya Belajar Auditorial

Gaya belajar auditorial adalah gaya belajar dengan cara mendengar. Individu dengan gaya belajar ini lebih dominan dalam menggunakan indera pendengaran untuk melakukan kegiatan belajar. Indvidu yang memiliki gaya belajar ini mudah belajar serta mudah menangkap stimulus (rangsangan) apabila melalui alat indera pendengaran (telinga). Individu dengan gaya belajar auditorial memiliki kekuatan pada kemampuannya untuk mendengar. Ciri-ciri individu yang memiliki tipe gaya belajar auditorial yaitu [3]:

1. Saat bekerja sering bicara pada diri sendiri.
2. Mudah terganggu oleh keributan atau hiruk pikuk disekitarnya.
3. Sering menggerakkan bibir dan mengucapkan tulisan dibuku ketika membaca.
4. Senang membaca dengan keras dan mendengarkan sesuatu.
5. Dapat mengulangi kembali dan menirukan nada, birama, dan warna suara dengan mudah.
6. Merasa kesulitan untuk menulis tetapi mudah dalam bercerita.
7. Pembicara yang fasih.
8. Lebih suka *music* daripada seni yang lainnya.

9. Lebih mudah belajar dengan mendengarkan dan mengingat apa yang didiskusikan daripada yang dilihat.
10. Suka berbicara, berdiskusi, dan menjelaskan sesuatu dengan panjang lebar.
11. Lebih pandai mengeja dengan keras daripada menuliskannya.

2.2.3.3 Gaya Belajar Kinestetik

Gaya belajar kinstetik adalah gaya belajar dengan cara bergerak, bekerja, dan menyentuh. Maksudnya gaya belajar dengan mengutamakan indera perasa dan gerakan-gerakan fisik. Individu dengan gaya belajar ini lebih mudah menangkap pelajaran apabila bergerak, meraba atau mengambil tindakan. Ciri-ciri individu yang memiliki tipe gaya belajar kinestetik yaitu [3]:

1. Berbicara dengan perlahan.
2. Menyentuh untuk mendapatkan perhatian.
3. Berdiri dekat ketika berbicara dengan orang.
4. Selalu berorientasi dengan fisik dan banyak bergerak.
5. Menghafal dengan cara berjalan dan melihat.
6. Menggunakan jari sebagai penunjuk ketika membaca.
7. Banyak menggunakan isyarat tubuh.
8. Tidak dapat duduk diam untuk waktu lama.
9. Memungkinkan tulisannya jelek.
10. Ingin melakukan segala sesuatu.
11. Menyukai permainan yang menyibukkan.

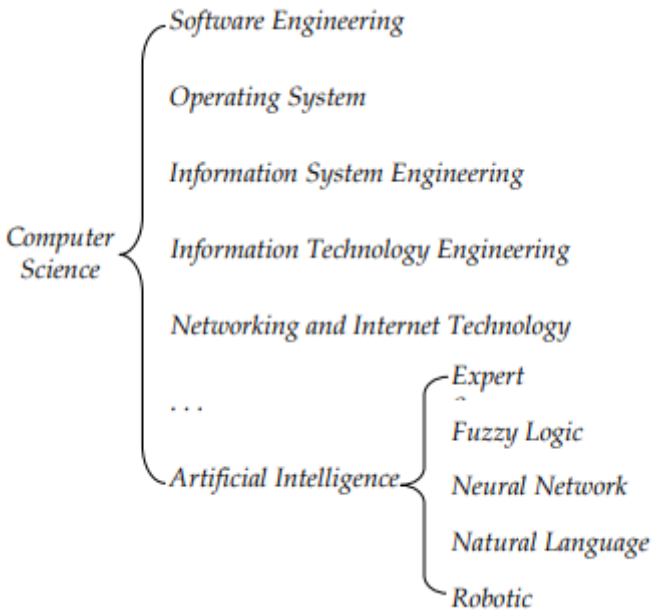
2.2.4 *Artificial Intelligence* (AI)

Kecerdasan buatan atau *artificial intelligence* (AI), definisinya menurut para beberapa pakar [12]:

1. Schalkoff (1990), AI adalah bidang studi yang berusaha menerangkan dan meniru perilaku cerdas dalam bentuk proses komputasi.
2. Rich dan Knight (1991), AI adalah studi tengtang cara membuat computer melakukan sesuatu yang sampai saat ini orang dapat melakukannya lebih baik.
3. Luger dan Stubblefield (1993), AI adalah cabang ilmu computer yang berhubungan dengan otomasi perilaku yang cerdas.

4. Hang dan Keen (1996), AI adalah bidang studi yang berhubungan dengan penangkapan, pemodelan, dan penyimpanan kecerdasan manusia dalam sebuah sistem teknologi informasi sehingga sistem tersebut dapat memfasilitasi proses pengambilan keputusan yang biasanya dilakukan oleh manusia.

Kecerdasan berasal dari kata dasar cerdas. Cerdas dapat memiliki konotasi makna lebih baik, cepat, *capable*, *adapted* dengan kondisi umumnya atau normal. Cerdas juga dapat berarti kemampuan untuk mengerti atau memahami. Kecerdasan adalah kemampuan manusia untuk memperoleh pengetahuan dan pandai melaksanakannya dalam praktek. Kecerdasan buatan merupakan sub bidang ilmu *computer* yang khusus ditujukan untuk membuat perangkat lunak dan perangkat keras yang sepenuhnya bisa menirukan bisa menirukan beberapa fungsi otak manusia atau cabang ilmu *computer* yang mempelajari otomasisasi tingkah laku cerdas (*intelligence*) [12].



Gambar 2. 2 Bagan Kedudukan Ilmu Kecerdasan Buatan *[12]*

Kecerdasan harus didasarkan pada prinsip-prinsip teoritikal dan terapan yang menyangkut [12]:

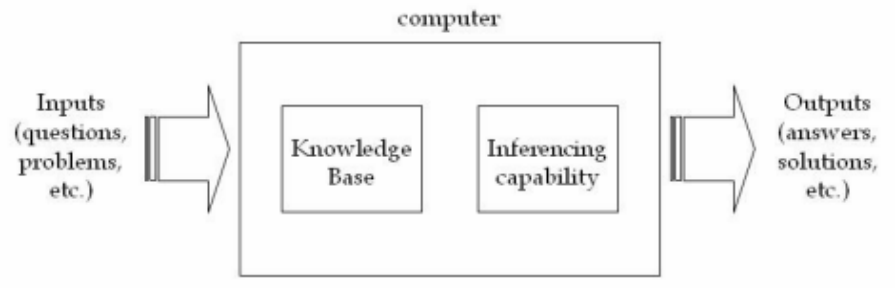
a. Struktur data yang digunakan dalam representasi pengetahuan (*knowledge representation*)
b. Algoritma yang diperlukan dalam penerapan pengetahuan itu,
c. Teknik-teknik bahasa dan pemrograman uang dipakai dalam implementasinya.

Kecerdasan buatan menawarkan baik media atau uji teori kecerdasan. Teori-teori ini dapat dinyatakan dalam bahasa program computer dan dibuktikan melalui eksekusinya pada *computer*.

Bagian utama aplikasi kecerdasan buatan adalah pengetahuan (*knowledge*), yaitu suatu pengertian tentang beberapa wilayah subyek yang diperoleh melalui pendidikan dan pengalaman. Bagian yang dibutuhkan untuk aplikasi kecerdasan buatan [12]:

a. Basis pengetahuan (*knowledge base*)
b. Motor inferensi (*inference engine*)

Pengetahuan merupakan informasi terorganisir dan teranalisis agar bisa lebih mudah dimengerti dan bisa diterapkan pada pemecahan masalah dan pengambilan keputusan. Pengetahuan terdiri atas fakta, pemikiran teori, prosedur, dan hubungannya satu sama lain [12].



Gambar 2. 3 Penerapan Konsep Kecerdasan Buatan dalam komputer *[12]*

Bidang-bidang teknik kecerdasan buatan diantaranya adalah [12]:

1. Sistem pakar (*expert system*)
2. Pengolahan bahasa alami (*natural language processing*)
3. Pengenalan ucapan (*speech recognition*)
4. Pengolahan citra (*image processing*)
5. Robotik (*robotics*) dan sensor
6. Logika samar (*fuzzy logic*)
7. Algoritma genetika (*genetic algorithm*)
8. Jaringan saraf tiruan (*neural netwoks*)
9. *Intelligence computer-aided instruction*
10. *Game playing*
11. *Evolutionary computing* (optimasi)

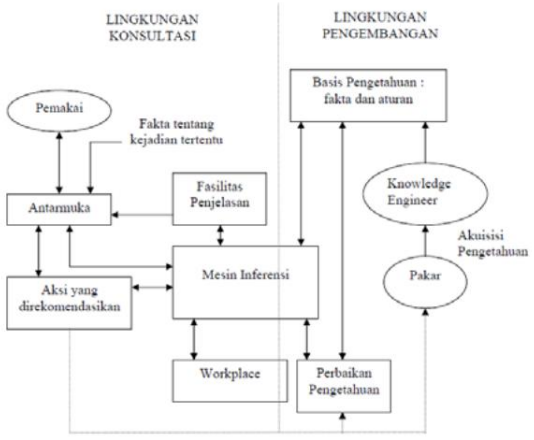
12. *Probabilistic reasoning* (mengakomodai ketidakpastian)

2.2.5 Sistem Pakar

Sistem pakar adalah sebuah program computer yang dirancang untuk memodelkan masalah seperti layaknya seorang pakar (*human expert*). Sistem pakar yang baik dirancang agar dapat menyelesaikan suatu permasalahan tertentu dengan meniru kerja dari para ahli di bidangnya masing-masing. Sistem pakar terkadang lebih baik unjuk kerjanya daripada seorang pakar manusia, sehingga sistem pakar dapat disimpulkan yaitu sebagai sebuah kepakaran yang ditransfer dari sebuah pakar (atau sumber kepakaran yang lain) ke *computer*, pengetahuan yang ada kemudian disimpan dalam memori computer dan pengguna dapat berkonsultasi dengan computer untuk suatu keperluan tertentu, lalu computer dapat menyimpulkan seperti layaknya seorang pakar, kemudian menjelaskannya kepada pengguna tersebut, dengan menyertakan alasan-alasannya [12].

Definisi sistem pakar [12]:

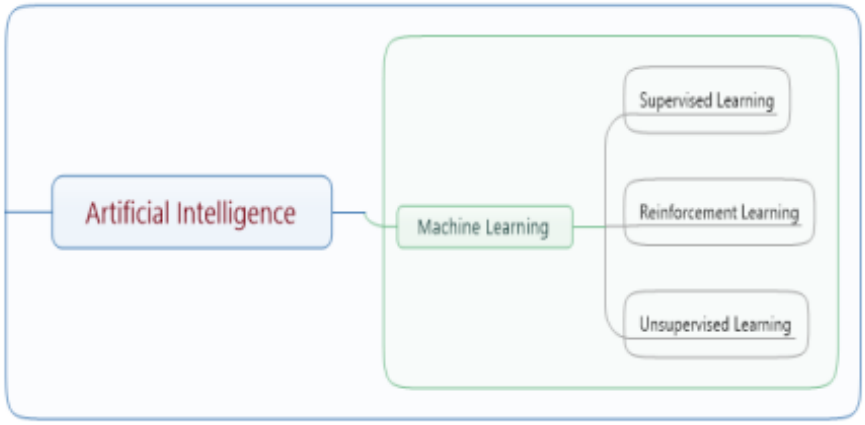
1. Menurut Durkin, sistem pakar merupakan suatu program *computer* yang dirancang untuk memodelkan kemampuan penyelesaian masalah yang dilakukan oleh seorang pakar.
2. Menurut Ignizio, sistem pakar merupakan suatu model dan prosedur yang berkaitan dalam suatu domain tertentu yang mana tingkat keahliannya dapat dibandingkan dengan keahlian seorang pakar.
3. Menurut Giarratano dan Riley, sistem pakar merupakan suatu sistem *computer* yang bisa menyamai atau meniru kemampuan seorang pakar.



Gambar 2. 4 Struktur Sistem Pakar *[12]*

2.2.6 *Machine Learning*

Machine learning atau dikenal dengan pembelajaran mesin adalah ilmu *computer* yang bisa bekerja tanpa diprogram secara eksplisit. *Machine learning* bergantung pada pola dan kesimpulan. Untuk mendapatkan pola dan kesimpulan tersebut, algoritma *machine learning* menghasilkan model matematika yang didasari dari data sampel yang sering disebut dengan '*training data*'. *Machine learning* adalah bidang keilmuan yang mempelajari bagaimana membuat program yang dapat menghasilkan pengetahuan baru dari pengetahuan yang sudah ada (disebut *experience* atau data) di luar pengetahuan yang "diprogram" secara langsung pada program [13].



Gambar 2. 5 Skema Artificial Intelligence dan Machine Learning *[14]*

Machine learning terbagi menjadi tiga kategori yaitu *supervised learning*, *unsupervised learning*, dan *reinforcement learning* [14].

1. *Supervised learning*

Supervised learning adalah metode klasifikasi dimana kumpulan data sepenuhnya diberikan label untuk mengklasifikasikan kelas yang tidak dikenal. *Supervised learning* dikelompokkan lebih lanjut dalam masalah klasifikasi dan regresi. Masalah klasifikasi adalah ketika variable *output* berbentuk kategori, seperti merah atau biru atau penyakit dan tidak ada penyakit. Sedangkan masalah regresi adalah ketika variabel *output* adalah nilai riil, seperti dollar atau berat. *Supervised learning* memiliki beberapa algoritma popular seperti *Back-propagation, Linear Regression, Random Forest, Support Vector Machines, Naïve*

Bayesian, Metode Rocchio, Decision Tree, k-Nearest Neighbor, Neural Network, Logistic Regression, dan *Neural Network*. [14]

2. *Unsupervised learning*

Unsupervised learning sering disebut cluster dikarenakan tidak ada kebutuhan untuk pemberian label dalam kumpulan data dan hasilnya tidak mengidentifikasi contoh di kelas yang telah ditentukan. Dalam pembelajaran *unsupervised learning*, sistem disediakan dengan beberapa *input* sampel tetapi tidak ada output yang hadir. Karena tidak ada *output* yang diinginkan disini kategorisasi dilakukan sehingga algoritma membedakan dengan benar antara kumpulan data. *Unsupervised learning* dikelompokkan lebih lanjur dalam masalah *clustering* dan asosiasi. Masalah pengelompokkan (*clustering*) adalah tempat untuk menemukan pengelompokkan yang melekat dalam data, seperti mengelompokkan pelanggan berdasarkan pada perilaku pembelian. Sedangkan masalah asosiasi adalah aturan yang menggambarkan sebagian besar data yang ada, seperti orang yang membeli A juga cenderung membeli B. *Unsupervised learning* memiliki beberapa algoritma popular seperti *k-means, Apriori, Independent Subspace Analysis* (ISA) [14].

3. *Reinforcement learning*

Reinforcement learning biasanya berada antara *supervised learning* dan *unsupervised learning*, teknik ini bekerja dalam lingkungan yang dinamis dimana komsepnya harus menyelesaikan tujuan tanpa adanya pemberitahuan dari computer secara eksplisit jika tujuan tersebut telah tercapai. *Reinforcement learning* berasal dari teori belajar hewan. Pembelajaran ini tidak memerlukan pengetahuan sebelumnya, dapat secara mandiri mendapatkan kebijakan opsional dengan pengetahuan yang diperoleh melalui coba-coba dan terus berinteraksi dengan lingkungan yang dinamis. Masalah *reinforcement learning* diselesaikan dengan mempelajari pengalaman baru melalui *trial-and-error*. Algoritma *reinforcement learning* terkait dengan algoritma pemrograan dinamis yang sering digunakan untuk menyelesaikan masalah optimisasi [14].

2.2.7 *Data Mining*

Data mining adalah kegiatan menemukan pola yang menarik dari data jumlah besar. Data tersebut dapat disimpan dalam database, data warehouse atau

penyimpanan infomasi lainnya. *Data mining* berkaitan dengan bidang ilmu-ilmu lain seperti *database system*, *data warehousing*, *statistic*, *machine learning*, information retrieval, dan komputasi tingkat tinggi [15].

Data mining merupakan bagian dari *Knowledge Discovery in Database* (KDD) yang merupakan proses ekstrasi informasi yang berguna, tidak diketahui sebelumnya, dan tersembunyi dari data [16]. KDD adalah kegiatan yang meliputi pengumpulan, pemakaian data, historis untuk menemukan keteraturan, pola atau hubungan dalam set data berukuran besar [17].

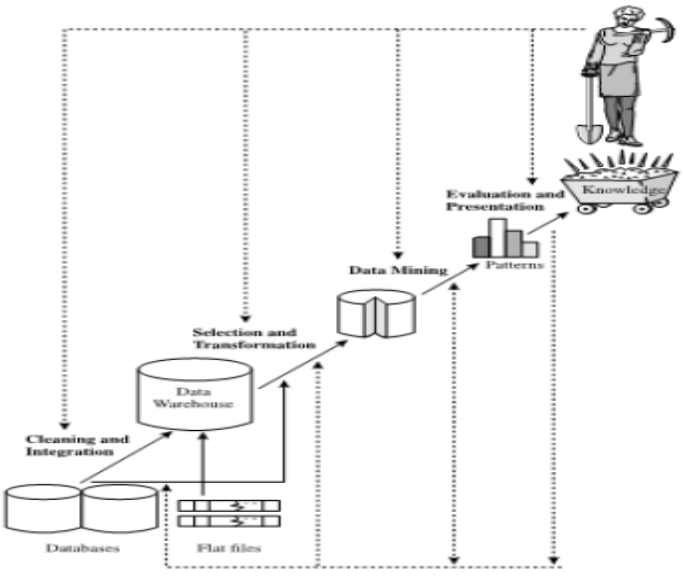
1) Metode pelatihan

Secara garis besar metode pelatihan dalam teknik-teknik *data mining* dibedakan ke dalam dua pendekatan, yaitu [17]:

a. *Unsupervised learning,* metode ini diterapkan tanpa adanya latihan (*training*) dan tanpa ada guru (*teacher*). Guru disini adalah label dari data.
b. *Supervised learning*, yaitu metode belajar dengan adanya latihan dan pelatih. Dalam pendekatan ini, untuk menemukan fungsi keputusan, fungsi pemisah atau fungsi regresi digunakan beberapa contoh data yang mempunyai *output* atau label selama proses *training*.

2) Tahap-tahap *data mining*

Sebagai suatu rangkaian proses, data mining dapat dibagi menjadi bebrapa tahap proses seperti dilustrasikan pada gambar 2.6. Tahap-tahap tersebut bersifat interaktif, pemakai terlibat langsung atau dengan perantaraan *knowledge base*.



Gambar 2. 6 Tahap-tahap *Data Mining [17]*

Tahap-tahap dalam data mining adalah sebagai berikut [17]:

1. Pembersihan data (*data cleaning*)

Pembersihan data merupakan proses menghilangkan *noise* dan data yang tidak konsisten atau tidak relevan.

2. Integrasi data (*data integration*)

Integrasi data merupakan penggabungan data dari berbagai *database* ke dalam duatu *database* baru.

3. Seleksi data (*data selection*)

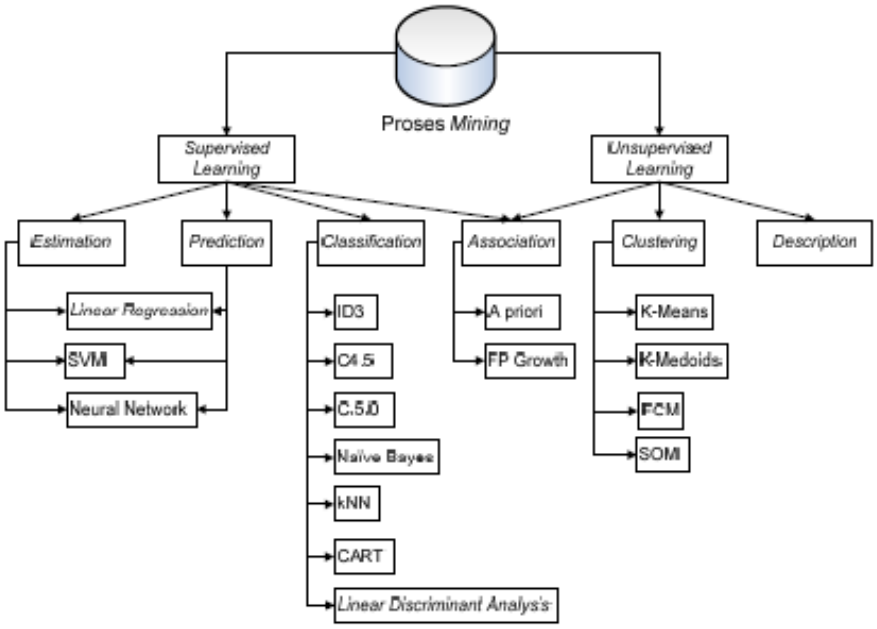
Data yang ada pada database seringkali tidak semuanya dipakai, oleh karena itu hanya data yang sesuai untuk dianalisis yang akan diambil dari *database*.

4. Transformasi data (*data transformation*)

Data diubah atau digabung ke dalam format yang sesuai untuk diproses dalam *data mining*.

5. Proses *mining*

Merupakan proses utama saat metode diterapkan untuk menemukan pengetahuan berharga dan tersembunyi dari data. Beberapa metode yang dapat digunakan berdasarkan pengelompokan data *mining* dapat dilihat pada gambar 2.3.



Gambar 2. 7 Beberapa Metode *Data Mining [17]*

6. Evaluasi pola (*pattern evaluation*)

Untuk mengidentifikasi pola-pola menarik ke dalam *knowledge based* yang ditemukan.

7. Presentasi pengetahuan (*knowledge presentation*)

Merupakan visualisasi dan penyajian pengetahuan mengenai metode yang digunakan untuk memperoleh pengetahuan yang diperoleh pengguna.

Peranan *data mining* menurut Daniel T. Larose (2005) [18]:

1. Deskripsi

Menemukan jalan untuk mendeskripsikan pola dan kecenderungan dalam data.

2. Estimasi

Mirip dengan klasifikasi, tapi nilai *attribute* target berupa numerik bukan kategori.

3. Prediksi

Prediksi mirip dengan klasifikasi dan estimasi, namun untuk prediksi terdapat kesalahan hasil ke depannya.

4. Klasifikasi

Adanya variabel target bertipe kategori. Model *data mining* mengujikan sejumlah *record*, dan di setiap *record* berisi variabel target dan sekumpulan variable pemrediksi. Misalnya:

a) Menentukan apakah terjadi kecurangan transaksi *credit card*.
b) Penempatan murid baru jalur tertentu berdasarkan kebutuhan.
c) Mendiagnosis jenis penyakit yang muncul.

5. Clustering

Merupakan pengelompokan dalam suatu grup yagn memiliki tingkat kesamaan tinggi dan sebaliknya memiliki perbedaan tinggi (kesamaan yang rendah) terhadap kelompok berbeda.

6. Association

Menemukan artibut yang mana bersama-sama.

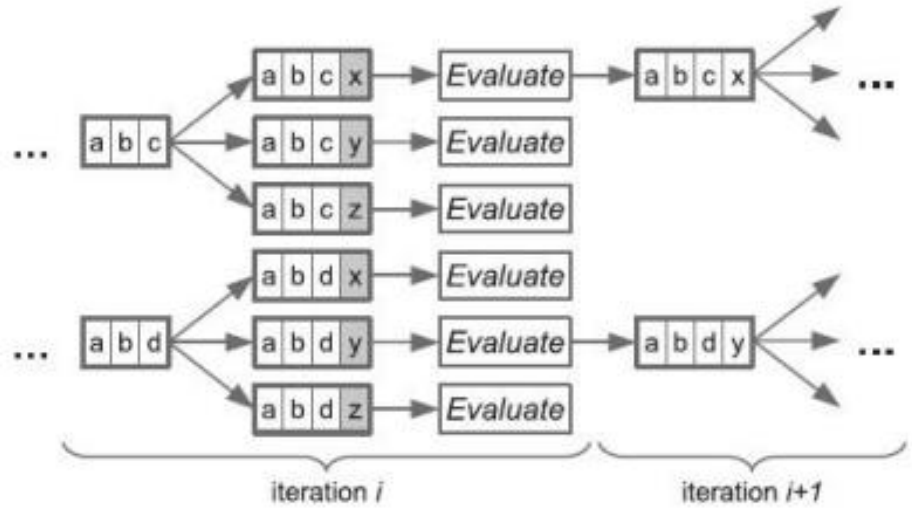
2.2.8 Seleksi Fitur

Seleksi fitur merupakan salah satu tahapan praproses yang berguna terutama dalam hal meningkatkan hasil akurasi, menghilangkan data yang tidak relevan, juga mengurangi dimensi data. Definisi seleksi fitur adalah mengamati sekumpulan fitur

kemudian dipilih beberapa fitur yang mampu memberikan hasil terbaik untuk klasifikasi [15].

2.2.9 *Forward Selection*

Forward selection adalah salah satu prosedur bertahap yang mempunyai tujuan untuk menambah variabel yang dikendalikan satu persatu ke dalam persamaan, *forward selection* dimulai dengan fitur himpunan kosong (nol) kemudian menambahkan fitur yang terpakai pada putaran pertama, sampai dengan semua fitur dievaluasi masing-masing. Salah satu fitur ditambahkan pada fitur himpunan yang merupakan bagian dari fitur sebelumnya dan fitur yang baru dibuat kemudian dievaluasi kembali. Untuk mengurangi jumlah evaluasi, hanya *subset* fitur terbaik yang akan disimpan seperti ditunjukkan gambar 2.8 [15].



Gambar 2. 8 Metode *Forward Selection [15]*

Data yang di *training* dilakukan secara tertahap yakni dimulai dari 1 variabel sampai pada tingkat atau jumlah variabel yang menghasilkan performa atau nilai akurasi yang paling baik atau *error* terkecil. Misalnya pada pengujian data dengan 2 variabel menghasilkan *error* lebih kecil dan ketika diujikan lagi dengan 3 variabel dan menghasilkan nilai *error* lebih besar dibandingkan dengan 2 variabel maka *error* terkecil didapatkan pada *variable* ke 2 yang berarti *variable* ke 2 signifikan, proses akan dihentikan jika semua *variable* independen sudah diujikan. Algoritma *forward selection* akan diujikan pada setiap data, yaitu mulai dari data dengan 1

variabel periode sampai pada data dengan 10 variabel periode yang kemudian dibandingkan data mana yang menghasilkan akurasi yang paling baik (Reif dan Shafait, 2014) [15].

Forward selection didasarkan pada model regresi linear, prosedur *forward selection* dapat dirumuskan sebagai berikut [19]:

1. Menentukan model awal

$$\hat{y} = b_0 \quad (2.1)$$

Memasukkan variable respon dengan setiap variabel berprediktor, misalnya $X1. X2, \ldots\ldots X_n$ yang terkait dengan $\hat{y}$. Misalkan X1 sehingga membentuk model:

$$\hat{y} = b_0 + b_1 X_1 \quad (2.2)$$

2. Uji F terhadap peubah pertama yang terpilih

Jika $Fhitung < Ftabel$ maka peubah terpilih dibuang dan proses dihentikan;

Apabila $Fhitung > Ftabel$ maka peubah terpilih memiliki pengaruh nyata terhadap peubah terkait y; sehingga layak untuk diperhitungkan di dalam model.

3. Masukan peubah bebas terpilih (yang paling signifikan) ke dalam model:

$$\hat{y} = b_0 + b_1 X_1 + b_2 X_2 \quad (2.3)$$

4. Uji F, jika $Fhitung < Ftabel$ maka proses dihentikan dan model terbaik adalah model sebelumnya;

Namun jika $Fhitung \geq Ftabel$, variable peubah bebas layak untuk dimasukan ke dalam model dan kembali ke langkah 3, proses akan berakhir jika tidak ada lagi peubah yang tersisa yang bisa dimasukkan ke dalam model.

2.2.10 *Naïve Bayes*

Naïve bayes adalah suatu metode yang dapat digunakan untuk memperkirakan atau memprediksi suatu *class* dari suatu objek yang kelasnya tidak diketahui dari masing-masing kelompok atribut yang ada, serta menentukan *class* yang paling optimal berdasarkan pengaruh yang didapat dari hasil pengamatan [15]. Klasifikasi-klasifikasi *bayes* adalah klasifikasi statistik yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu *class* (Lieng, et al, 2014) [15]. Prediksi bayes didasarkan pada formula teorema *Bayes* dengan formula umum *Bayes* sebagai berikut [15]:

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)} \quad (2.4)$$

Keterangan [15]:

B = Data dengan class yang belum diketahui

A = Hipotesis data B merupakan suatu class spesifik

$P(A|B)$ = Probabilitas hipotesis A berdasar kondisi B

$P(A)$ = Probabilitas hipotesis A

$P(B|A)$ = Probabilitas B berdasarkan kondisi pada hipotesis A

$P(B)$ = Probabilitas dari B

Naïve Bayes Classifier atau disebut juga dengan *Bayesian Classification* merupakan metode pengklasifikasian statistik yang didasarkan pada teorema *Bayes* yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu kelas. *Bayesian Classification* terbukti memiliki akurasi dan kecepatan yang tinggi saat diaplikasikan ke dalam *database* yang besar [20].

Peluang bersyarat artibut kategorikal dinyatakan dalam bentuk sebagai berikut [20]:

$$P(A_i|C_j) = \frac{|A_{ij}|}{N_{Cj}} \quad (2.5)$$

Keterangan, dimana:

$|A_{ij}|$ = jumlah contoh pelatihan dari kelas $|A_{ij}|$ yang menerima nilai C_j

Jika hasilnya adalah nol, maka menggunakan pendekatan berikut [20]:

$$P(A_i|C_j) = \frac{n_C + n_{equiv}\, p}{n + n_{equiv}} \quad (2.6)$$

Keterangan, dimana [20]:

n = total dari jumlah hasil dari kelas C_j

n_C = jumlah contoh pelatihan dari kelas A_i yang menerima nilai C_j

n_{equiv} = nilai konstan dari ukuran sampel yang ekuivalen

P = peluang estimasi *prior*, $P = {}^{1}/_{k}$ dimana *k* adalah jumlah kelas dalam variabel target.

Peluang bersyarat kontinu dinyatakan dalam bentuk berikut [20]:

$$P(A_i|C_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}}} exp\left[-\frac{(A_i - \mu_{ij})^2}{2(\sigma_{ij})^2}\right] \quad (2.7)$$

Parameter μ_{ij} dapat diestimasi berdasarkan sampel *mean* A_i untuk seluruh hasil pelatihan yang dimiliki kelas C_j. Dengan cara sama, $(\sigma_{ij})^2$ dapat diestimasi dari sampel varian (s^2) hasil dari pelatihan tersebut.

Prior probability atau $P(X)$ merupakan derajat *belive* atau sebuah informasi *probabilistic*. *Prior probability* berfngsi sebagai acuan *probabilistic* apabila tidak ada informasi lain yang bisa digunakan sebagai suatu kondisi [21].

$$P = \frac{X}{A} \tag{2.8}$$

Dimana:

P = Nilai *Prior*

X = Jumlah data pada tiap kelas

A = Jumlah data pada seluruh data

Conditional probability (CPT) merupakan sebuah tabel yang berisi sebuah kemungkinan *probabilistic* A dan B. *Conditional probability* sendiri adalah bagian dari komponen kuantitatif dari metode *bayesian network* (Putra, 2018). Apabila pada prosesnya informasi baru atau nilai baru telah didapatkan maka nilai atau informasi baru dari *probabilistic* tersebut yang digunakan sebagai acuan. *probabilistic* ini disebut sebagai kemungkinan bersyarat atau juga disebut *conditional probability* [21].

$$\mathrm{C}PT = \frac{P(A|B)}{P(B)} \tag{2.9}$$

Keterangan:

CPT = *Conditional Probability*

P$(A|B)$= *Probabilistic* terjadi dengan syarat kemungkinan B terjadi

$P(B)$ = Total banyaknya kemunculan data B

Normalizing constan atau juga bisa disebut normalisasi konstanta adalah suatu langkah yang digunakan pada metode *bayes* untuk melakukan normalisasi dari hasil *Conditional Probability*. Pada teorema *bayes* disebutkan bahwa perbandingan antara *Posterior Probability* sama dengan *Conditional Probability*. Dimana P(A) adalah nilai dari *Prior* dengan asumsi nilai dari *prior* adalah benar sedangkan P$(A|B)$ adalah nilai dari *Conditional Probability* dari data ataupun gejala yang diinputkan [21].

2.2.11 *Confusion Matrix*

Confusion matrix sebagai indikator analisis performa *classifier* dalam mengidentifikasi tupel (sampel) dari kelas yang berbeda. *Confusion matrix* juga dikenal dengan istilah *true positive* (tupel positif) dengan label benar, sedangkan *true negative* (tupel negatif) dengan label benar. Ada juga *false positive* yang merupakan tupel negatif dengan label salah, sedangkan *false negative* (tupel positif) dengan label salah [16].

Tabel 2. 1 Confusion Matrix 3 Kelas *[22]*

Confusion Matrix		Predicted		
		Class 1	Class 2	Class 3
Actual	Class 1	A	B	C
	Class 2	D	E	F
	Class 3	G	H	I

True positives True Negatives Misclassified cases.

Berikut ini merupakan formulasi perhitungan nilai akurasi, presisi dan *recall* [22]:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} = \frac{TP\ Total}{Dataset\ Total} \quad (2.10)$$

$$Precision = \frac{TP}{TP+FP} = \frac{TP}{Prediction\ Total} \quad (2.11)$$

$$Recall = \frac{TP}{TP+FN} = \frac{TP}{Actual\ Total} \quad (2.12)$$

Keterangan [22]:

1. *True Positive* (TP), label milik kelas dan diprediksi dengan benar.
2. *False Positive* (FP), label bukan milik kelas tetapi *classifier* diprediksi sebagai positif.
3. *False Negative* (FN), label bukan milik kelas dan diprediksi dengan benar.
4. *True Negative* (TN), label memang milik kelas tetapi diprediksi sebagai negatif.