BAB II

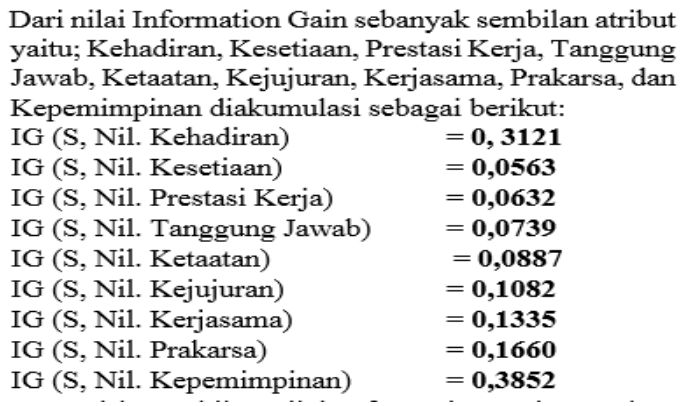
TINJAUAN PUSTAKA DAN LANDASAN TEORI

2.1. Tinjauan Pustaka

Tinjauan pustaka atau disebut juga kajian pustaka (*literature review*) merupakan sebuah aktivitas untuk meninjau atau mengkaji kembali berbagai literatur yang telah dipublikasikan oleh akademisi atau peneliti lain terkait topik yang diteliti. Peneliti telah merangkum kajian pustaka yang akan menjadi acuan peneliti melakukan penelitian, diantaranya:

Julce Adiana Sidette, Eko Sediyono, Oky Dwi Nurhayati (2014) melakukan penelitian dengan topik "Pendekatan Metode Pohon Keputusan Menggunakan *Decision Tree* Untuk Sistem Informasi Pengukuran Kinerja PNS". Mereka melakukan observasi dengan tahapan penelitian mengidentifikasi dan merumuskan masalah lalu diikuti dengan penentuan atribut yang menjadi objek penelitian.

Dimana 127 sampel data pegawai mulai dari P1 sampai P127 yang diambil penilaianya dan kelas yang mereka pilih adalah bagus dan buruk. Dari 127 pegawai yang masuk dalam kelas bagus adalah 56 dan yang masuk dalam kelas buruk adalah 71 dengan hasil *entropy* total 0,9899. lalu setelah *entropy* total diketahui dihitung *gain* dari tiap tiap atribut dan hasilnya dapat dilihat pada gambar dibawah ini.



Dari nilai Information Gain sebanyak sembilan atribut yaitu; Kehadiran, Kesetiaan, Prestasi Kerja, Tanggung Jawab, Ketaatan, Kejujuran, Kerjasama, Prakarsa, dan Kepemimpinan diakumulasi sebagai berikut:

IG (S, Nil. Kehadiran) = **0, 3121**
IG (S, Nil. Kesetiaan) = **0,0563**
IG (S, Nil. Prestasi Kerja) = **0,0632**
IG (S, Nil. Tanggung Jawab) = **0,0739**
IG (S, Nil. Ketaatan) = **0,0887**
IG (S, Nil. Kejujuran) = **0,1082**
IG (S, Nil. Kerjasama) = **0,1335**
IG (S, Nil. Prakarsa) = **0,1660**
IG (S, Nil. Kepemimpinan) = **0,3852**

Gambar 2.1 Informasi Nilai *Gain*

Dari hasil penelitian yang dilakukan Julce Adiana dkk maka dapat disimpulkan nilai *gain* kepemimpinan adalah menjadi yang terbesar jadi bisa dipastikan akan menjadi *root* pada pohon keputusan.

Windy Julianto, Rika Yunitarini, Mochammad Kautsar Sophan (2014) dari Universitas Trunojoyo Madura melakukan penelitian tentang "Algoritma C.45 Penilaian Untuk Kinerja Karyawan" dengan permasalahan minimnya tatap muka antara manager dan karyawan sehingga mereka membuat sistem pendukung keputusan untuk menilai kinerja karyawan dengan metode *data mining.* Sistem yang mereka bangun memakai beberapa kriteria antara lain: komunikasi, orientasi prestasi, inisiatif, pemikiran analitis, kepedulian terhadap tugas, kerja sama, pelayanan pelanggan, kerapian administrasi, pengaturan kerja, kemampuan teknis dan fungsionalitas.

Metodologi dari penelitian mereka menggunakan algoritma C.45 untuk membangun sebuah pohon keputusan atau *decision tree* dan untuk tes ukuran dari teori informasinya menggunakan *entropy* dan informasi *gain*. Perhitungan *gain* masih memiliki kekurangan. Salah satu kekurangan tersebut yakni pemilihan atribut yang tidak relevan sebagai pemartisi pada suatu simpul dan *gain ratio* merupakan normalisasi yang memperhitungkan *entropy* total.

Data digunakan berasal dari data karyawan Gajah Mada Lumajang dan terdiri dari total data karyawan yang sudah dikerjakan maupun pelamar dengan total 364 karyawan. Data *training* 192, data *testing* 152 dengan nilai yang dihitung memakai *confusion matrix* dan hasilnya *precision* 60%, *recall* 88.24%, *accuracy* 92 %, *error rate* 7,96%.

Teguh Budi Santoso (2014) menganalisa prediksi loyalitas pelanggan dengan menerapkan metode C.45 dari Teknik Informatika, Universitas Satya Negara Indonesia. Data sampel diambil pada bulan oktober sampai dengan November 2013 dengan atribut usia, pelayanan, promosi, harga, citra perusahaan dan kepercayaan. Jumlah data 40 pelanggan, dalam perhitungan *entropy* dan *gain* hasil klasifikasi pada data sampel atribut pelayanan sebagai *root node* dengan nilai gain 0,7083 sedangkan yang lainya menjadi *child node.*

Dari ketiga jurnal rangkuman yang dapat diambil adalah teknik klasifikasi merupakan metode *data mining* yang menggolongkan atribut target dengan membangun sebuah pohon keputusan menggunakan sistem aplikasi atau *software open source* sebagai pengujianya dengan perhitungan *entropy* dan infomasi *gain* sebagi tes hasil teorinya serta algoritma C.45 sebagai modelnya.

Dari hasil rangkuman jurnal tersebut maka peneliti tertarik menggunakan metode yang sama yaitu menggunakan algoritma C.45 untuk membangun sebuah pohon keputusan atau *decision tree* dengan mencari informasi tersembunyi dari data dosen Fakultas Teknik Universitas Muhammadiyah Yogyakarta menggunakan atribut lama kerja.

2.2. Landasan Teori

Pada penelitian kuantitatif landasan teori sangat berperan penting dalam sebuah penelitian, karena tanpa landasan teori maka penelitian akan berujung pada kesalahan atau yang sering dikenal dengan istilah *trial and error.*

Setiap teori bisa dikatakan sebagai dugaan sementara, karena hal tersebut memerlukan pembuktian bahwa sebuah teori akan memperoleh arti penting mana kala lebih banyak melukiskan, menerangkan dan meramalkan gejala yang ada.

2.2.1. Data Mining

Data mining adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan mesin pembelajaran (*machine learning*) untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai *database* besar. Selain definisi di atas beberapa definisi juga diberikan seperti, "*data mining*

adalah serangkaian proses untuk menggali nilai tambah dari suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui secara manual." (Windy Julianto, 2014).

"*Data mining* merupakan analisis dari peninjauan kumpulan data untuk menemukan hubungan yang tidak diduga dan meringkas data dengan cara yang berbeda dengan sebelumnya, yang dapat dipahami dan bermanfaat bagi pemilik data". "*Data mining* adalah suatu proses menemukan hubungan yang berarti pola, dan kecenderungan dengan memeriksa dalam sekumpulan besar data yang tersimpan dengan menggunakan teknik pengenalan pola seperti teknik statistik dan matematika." (Teguh Budi Santoso, 2014).

Kemajuan luar biasa yang terus berlanjut dalam bidang *data mining* didorong oleh beberapa faktor, antara lain:

1. Pertumbuhan yang cepat dalam kumpulan data.
2. Penyimpanan data dalam *data warehouse*, sehingga seluruh perusahaan memiliki akses ke dalam *database* yang baik.
3. Adanya peningkatan akses data melalui navigasi web dan intranet.
4. Tekanan kompetisi bisnis untuk meningkatkan penguasaan pasar dalam globalisasi ekonomi.
5. Perkembangan yang hebat dalam kemampuan komputasi dan pengembangan kapasitas media penyimpanan.

Berdasarkan definisi yang telah disampaikan, hal penting yang terkait dengan *data mining* adalah:

1. *Data mining* merupakan suatu proses otomatis terhadap data yang sudah ada.
2. Data yang akan diproses berupa data yang sangat besar.
3. Tujuan *data mining* adalah mendapatkan hubungan atau pola yang mungkin memberikan indikasi yang bermanfaat.

Hubungan yang dicari dalam *data mining* dapat berupa hubungan antara dua atau lebih dalam satu dimensi. Misalnya dalam dimensi produk, dapat dilihat keterkaitan pembelian suatu produk dengan produk yang lain. Selain itu, hubungan juga dapat dilihat antara dua atau lebih atribut dan dua atau lebih objek.

Sementara itu, penemuan pola merupakan keluaran lain dari *data mining*. Misalkan sebuah perusahaan yang akan meningkatkan fasilitas kartu kredit dari pelanggan, maka perusahaan akan mencari pola dari pelanggan – pelanggan yang ada untuk mengetahui pelanggan yang potensial dan pelanggan yang tidak potensial.

Beberapa definisi awal dari *data mining* meyertakan fokus pada proses otomatisasi. Berry dan Linoff, (2004) dalam buku *Data Mining Technique for Marketing, Sales, and Customer Support* mendefinisikan *data mining* sebagai suatu proses eksplorasi dan analisis secara otomatis maupun semi otomatis terhadap data dalam jumlah besar dengan tujuan menemukan pola atau aturan yang berarti (Larose, 2006).

Pernyataan tersebut menegaskan bahwa dalam *data mining* otomatisasi tidak menggantikan campur tangan manusia. Manusia harus ikut aktif dalam setiap *fase* dalam proses *data mining.* Kehebatan kemampuan algoritma *data mining* yang terdapat dalam perangkat lunak analisis yang terdapat saat ini memungkinkan terjadinya kesalahan penggunaan yang berakibat fatal. Pengguna mungkin menerapkan analisis yang tidak tepat terhadap kumpulan data dengan menggunakan pendekatan yang berbeda. Oleh karenanya, dibutuhkan pemahaman tentang statistik dan struktur model matematika yang mendasari kerja perangkat lunak (Larose, 2006).

Data mining bukanlah suatu bidang yang sama sekali baru. Salah satu kesulitan untuk mendefinisikan *data mining* adalah kenyataan bahwa *data mining* mewarisi banyak aspek dan teknik dari bidang – bidang ilmu yang sudah mapan terlebih dahulu. Terdapat beberapa teknik *data mining* yang sering disebut – sebut dalam literatur. Namun ada 3 teknik data mining yang populer, yaitu:

1. *Association Rule Mining*

   Association Rule Mining (Asosiasi Peraturan Pertimbangan) adalah teknik *mining* untuk menemukan asosiatif antara kombinasi atribut. Contoh dari aturan asosiatif dari analisa pembelian disuatu pasar swalayan dapat mengatur penempatan barangnya atau merancang strategi pemasaran dengan memakai kupon diskon untuk kombinasi barang tertentu.

2. *Clustering* (Pengelompokkan)

   Berbeda dengan *association rule mining* dan klasifikasi dimana kelas data telah ditentukan sebelumnya, pengklusteran dapat dipakai untuk memberikan *label* pada kelas data yang belum diketahui. Karena itu pengklusteran sering digolongkan sebagai metode *unsupervised learning*. Prinsip pengklusteran adalah memaksimalkan kesamaan antar kluster. Pengklusteran dapat dilakukan pada data yang memiliki beberapa atribut yang dipetakan sebagai ruang multidimensi.

3. Klasifikasi

   Dalam klasifikasi, terdapat target variabel kategori. Sebagai contoh, penggolongan pendapatan dapat dipisahkan dalam tiga kategori, yaitu pendapatan tinggi, pendapatan sedang, pendapatan rendah.

Data – data yang ada, tidak dapat langsung diolah dengan menggunakan sistem *data mining*. Data – data tersebut harus dipersiapkan lebih dulu agar hasil yang diperoleh lebih maksimal, dan waktu komputasinya lebih minimal. Proses persiapan data ini sendiri dapat mencapai 60% dari keseluruhan proses dalam *data mining*.

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 KDD adalah *data mining.* Proses KDD secara garis besar dapat dijelaskan sebagai berikut.

1. *Data Selection*

Pemilihan (seleksi) data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam KDD dimulai. Data hasil seleksi yang akan digunakan untuk proses *data mining,* disimpan dalam suatu berkas, terpisah dari basis data operasional.

2. *Pre-processing/Cleaning*

Sebelum proses *data mining* dapat dilaksanakan, perlu dilakukan proses *cleaning* pada data yang menjadi fokus KDD. Proses *cleaning* mencakup antara lain membuang

duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data, seperti kesalahan cetak (tipografi). Juga dilakukan proses *enrichment*, yaitu proses "memperkaya" data yang sudah ada dengan data atau informasi lain yang relevan dan diperlukan untuk KDD, seperti data atau informasi eksternal.

3. *Transformation*

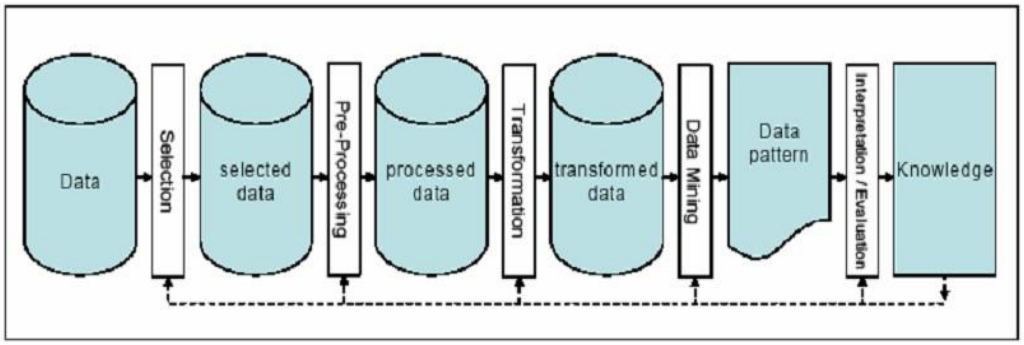
Coding adalah proses transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses *data mining*. Proses *coding* dalam KDD merupakan proses kreatif dan sangat tergantung pada jenis atau pola informasi yang akan dicari dalam basis data.

4. *Data mining*

Data mining adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik, metode, atau algoritma dalam *data mining* sangat bervariasi. Pemilihan metode dan algoritma yang tepat sangat bergantung pada tujuan dan proses KDD secara keseluruhan.

5. *Interpretation/Evalution*

Pola informasi yang dihasilkan dari proses *data mining* perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan. Tahap ini merupakan bagian dari proses KDD yang disebut *interpretation*. Tahap ini mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesis yang ada sebelumnya. Penjelasan di atas dapat direfresentasikan pada gambar dibawah ini.



Gambar 2.2 Proses dari KDD

2.2.2. Klasifikasi

Metode klasifikasi adalah sebuah metode dari *data mining* yang digunakan untuk memprediksi kategori atau kelas dari suatu *data instance* berdasarkan sekumpulan atribut-atribut dari data tersebut. Atribut yang digunakan mungkin bersifat *categorical* (misalnya golongan darah: "A", "B", "O", dst), *ordinal* (misalnya urutan: *small, medium,* dan *large*), *integer-valued* (misalnya banyaknya suatu kata pada suatu paragraf), atau *real-valued* (misalnya suhu).

Kebanyakan algoritma yang menggunakan metode klasifikasi ini hanya menggunakan data yang bersifat diskret dan untuk data yang bersifat kontinu (*real-valued* dan *integer-valued*) maka data tersebut harus dijadikan diskret dengan cara memberikan *threshold* (misal lebih kecil dari 5 atau lebih besar dari 10) supaya data dapat terbagi menjadi grup-grup. Sebagai contoh dari metode klasifikasi adalah menentukan *e-mail* yang masuk termasuk kategori *spam* atau bukan *spam* atau menentukan diagnosis dari pasien berdasarkan umur, jenis kelamin, tekanan darah, dan sebagainya (Tan, 2004).

2.2.3. Algoritma ID3

Iterative Dichotomicer 3 (ID3) adalah algoritma *decision tree learning* (algoritma pembelajaran pohon keputusan) yang paling dasar. Algoritma ini melakukan pencarian secara rakus pada semua kemungkinan pohon keputusan.

Algoritma ID3 dapat diimplementasikan menggunakan fungsi rekursif (fungsi yang memanggil dirinya sendiri). Algoritma ID3 berusaha membangun *decision tree* (pohon keputusan) secara *top-down* (dari atas ke bawah), mulai dengan pertanyaan: "atribut mana yang pertama kali harus dicek dan diletakkan pada *root*?" pertanyaan ini dijawab dengan mengevaluasi semua atribut yang ada dengan menggunakan suatu ukuran statistik (yang banyak digunakan adalah *information gain*) untuk mengukur efektivitas suatu atribut dalam mengklasifikasikan kumpulan sampel data.

Karakteristik ID3 dalam membangun pohon keputusan adalah secara *top-down* dan *divide-and-conquer*. *Top-down* artinya pohon keputusan dibangun dari simpul akar ke daun, sementara *divide-and-conquer* artinya *training data* secara rekursif dipartisi kedalam bagian-bagian yang lebih kecil saat pembangunan pohon.

Decision Tree adalah sebuah struktur pohon, dimana setiap *node* pohon merepresentasikan atribut yang telah diuji, setiap cabang merupakan suatu pembagian hasil uji, dan *node* daun (*leaf*) merepresentasikan kelompok kelas tertentu. *Level node* teratas dari sebuah *decision tree* adalah *node* akar (*root*) yang biasanya berupa atribut yang paling memiliki pengaruh terbesar pada suatu kelas tertentu. Pada umumnya *decision tree* melakukan strategi pencarian secara *top-down* untuk solusinya. Pada proses mengklasifikasi data yang tidak diketahui, nilai atribut akan diuji dengan cara melacak jalur dari *node* akar (*root*) sampai *node* akhir (daun) dan kemudian akan diprediksi kelas yang dimiliki oleh suatu data baru tertentu.

2.2.4. Algoritma C.45

Algoritma C4.5 dan *decision tree* merupakan dua model yang tak terpisahkan, karena untuk membangun sebuah *decision tree*, dibutuhan algoritma C4.5. Di akhir tahun 1970 hingga di awal tahun 1980-an, J. Ross Quinlan seorang peneliti di bidang mesin pembelajaran mengembangkan sebuah model *decision tree* yang dinamakan ID3 (*Iterative Dichotomiser*), walaupun sebenarnya proyek ini telah dibuat sebelumnya oleh E.B. Hunt, J. Marin, dan P.T. Stone. Kemudian Quinlan membuat algoritma dari pengembangan ID3 yang dinamakan C4.5 yang berbasis *supervised learning*.

Serangkaian perbaikan yang dilakukan pada ID3 mencapai puncaknya dengan menghasilkan sebuah sistem praktis dan berpengaruh untuk *decision tree* yaitu C4.5. Perbaikan ini meliputi metode untuk menangani *numeric attributes, missing values, noisy* data, dan aturan yang menghasilkan *rules* dari *trees*.

Ada beberapa tahapan dalam membuat sebuah *decision tree* dalam algoritma C4.5 (Larose, 2005) yaitu :

1. Mempersiapkan data *training*. Data *training* biasanya diambil dari data histori yang pernah terjadi sebelumnya atau disebut data masa lalu dan sudah dikelompokkan dalam kelas – kelas tertentu.
2. Menghitung akar dari pohon. Akar akan diambil dari atribut yang akan terpilih, dengan cara menghitung nilai *gain* dari masing – masing atribut, nilai *gain* yang paling tinggi

yang akan menjadi akar pertama. Sebelum menghitung nilai *gain* dari atribut, hitung dahulu nilai *entropy*.

Untuk menghitung nilai *entropy* menggunakan persamaan I:

$$Entropy\ (S) = \sum_{i=1}^{n} - pi\ \log_2 pi$$

Keterangan:

S : himpunan kasus

A : fitur

n : jumlah partisi S

P_i : proporsi dari S_i terhadap S

Sementara itu perhitungan nilai *gain* menggunakan persamaan II:

$$Gain\ (S,A) = entropy\ (S) - \sum_{i=1}^{n} \frac{|Si|}{S} * Entropy\ (Si)$$

Keterangan:

S : himpunan kasus

A : atribut

n : jumlah partisi atribut A

|Si| : jumlah kasus pada partisi ke-i

|S| : jumlah kasus dalam S

Decision tree merupakan metode klasifikasi dan prediksi yang sangat kuat dan terkenal. Metode *decision tree* mengubah fakta yang sangat besar menjadi sebuah pohon keputusan yang mempresentasikan aturan. Aturan dapat dengan mudah dengan bahasa alami dan mereka juga dapat diekspresikan dalam bentuk basis data seperti *Structured Query Language* untuk mencari *record* pada kategori tertentu. *decision tree* juga berguna untuk memadukan antara eksplorasi data dan pemodelan, dia sangat bagus sebagai langkah awal dalam proses pemodelan bahkan ketika dijadikan sebagai model akhir dari beberapa teknik yang lain.

decision tree adalah sebuah struktur yang dapat digunakan untuk membagi kumpulan data yang besar menjadi himpunan – himpunan *record* yang lebih kecil dengan menerapkan serangkaian aturan keputusan. Sebuah model pohon keputusan terdiri dari sekumpulan aturan untuk membagi sejumlah populasi yang heterogen menjadi lebih kecil, lebih homogen dengan memperhatikan pada variabel tujuanya. Variabel tujuan biasanya dikelompokan dengan pasti dan model *decision tree* lebih mengarah pada perhitungan probabilitas dari tiap – tiap *record* terhadap kategori tersebut atau untuk mengklasifikasi *record* dengan mengelompokanya 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. Misalkan untuk menentukan main tenis, kriteria yang diperhatikan adalah cuaca, angin dan temperatur. Salah satu atribut merupakan atribut yang menyatakan data solusi per *item* data yang disebut target atribut. Atribut memiliki nilai – nilai yang dinamakan dengan *instance*. Misalkan atribut cuaca mempunyai *instance* berupa cerah, berawan dan hujan. Proses pada *decision tree* adalah mengubah bentuk data (tabel) menjadi model pohon, mengubah model pohon menjadi *rule* dan menyederhanakan *rule.*