BAB 2
TINJAUAN PUSTAKA

2.1 Tinjauan Pustaka

Sebelum melakukan penelitian adapun penulis mencari penelitian penelitian yang memungkinkan terkait dengan penelitian antara lain :

1. Analisis Kinerja Data Mining Algoritma C4.5 Dalam Menentukan Tingkat Minat Siswa yang Mendaftar di Kampus ABC (Yudhi Andrian, M. Rhifky Wayahdi, 2014) [1].

   Di dalam jurnal tersebut membahas minat siswa yang masuk ke Universitas masih menjadi hal yang perlu diketahui seberapa besar minat siswa tersebut mendaftar ke kampus ABC. Dalam hal ini dapat ditentukan dengan ilmu data mining. Data mining sendiri memiliki banyak algoritma. Dalam jurnal tersebut untuk membuat pohon keputusan, menggunakan algoritma C4.5. Sedangkan alur proses algoritma C4.5 untuk membangun pohon keputusan diperlukan pemilihan atribut sebagai akar, membuat cabang untuk tiap tiap nilai, membagi kasus dalam cabang, dan mengulangi khasus cabang hingga mencapai kelas yang sama. Dengan adanya algoritma C4.5 data tersebut diolah dengan mengambil data atribut atribut sederhana yaitu data siswa yang mendaftar dan tidak mendaftar pada kampus ABC yang diambil secara acak/random sebanyak 50 data sampel. Penggunaan algoritma C4.5 diterapkan untuk membangun pohon keputusan atau decision tree untuk memprediksi minat mahasiswa pada kampus ABC yang lebih baik sehingga menghasilkan dari 50 data diuji terdapat 40 siswa yang mendaftar dan 10 siswa yang tidak mendaftar dengan penentuan gain tertinggi dari atribut tahun tamat.

2. Analisis Dan Penerapan Algoritma C4.5 Dalam Data Mining Untuk Memprediksi Masa Studi Mahasiswa Berdasarkan Data Nilai Akademik (Selvia Lorena Br Ginting, Wendi Zarman, Ida Hamidah, 2014) [2].

   Dalam jurnal ini menjelaskan jurusan teknik komputer di Universitas Komputer Indonesia membutuhkan 8 semester atau 144 sks untuk lulus namun dari Jurusan Teknik Komputer Program Sarjana (S1) pada tahun 2007 hanya 1 orang yang lulus tepat waktu dalam 8 semester. Maka dari itu peneliti tersebut melakukan penelitian dengan membutuhkan data akademik mahasiswa yang sudah lulus atau dijadikan sebagai data training dan mahasiswa yang belum lulus yang berguna memprediksi ketepatan lulus mahasiswa. Dengan demikian dibutuhkan teknik data mining untuk memperoleh pohon keputusan yang dibangun dari algoritma C4.5. Atribut untuk menerapkan C4.5 menggunakan 3 sampel mata kuliah yaitu Algoritma Pemrograman, Fisika 1, Fisika 2 dan diambil gain tertinggi yaitu Fisika 2, dengan demikian Fisika 2 menjadi root atau node untuk membuat pohon keputusan atau decision tree. Dengan aplikasi data mining memprediksi masa studi mahasiswa dapat dilihat bahwa semakin banyaknya data training tingkat kecocokan semakin kecil dengan data testing. Dengan data training 112 tingkat kecocokan yang sangat kecil daripada data training dengan data 70.

3. Penerapan Algoritma C4.5 Untuk Penentuan Jurusan Mahasiswa (Liliana Swastina, 2013) [3].

   Dalam jurnal ini menjelaskan bagaimana menentukan jurusan yang tepat pada siswa SMU kelas XII. Dengan adanya data mining penentuan jurusan dapat dilakukan secara tepat. Algoritma yang dipakai disini yaitu perbandingan algoritma C4.5 dengan Naïve Bayes dengan membangun pohon keputusan atau decision tree sebagai penentu jurusan calon mahasiswa baru dengan latar belakang minat, dan kemampuannya sendiri. Objek disini adalah sekolah STMK Indonesia Banjarmasin dijadikan sebagai data untuk pengujian model dengan rapid miner sebagai simulasinya dengan memerlukan atribut Nama, Jenis Kelamin, Umur, Asal Sekolah, Jurusan Asal Sekolah, Nilai UAN, IPK

Semester 1, IPK Semester 2. Dari hasil pengujiannya ditemukan algoritma decision tree C4.5 memiliki tingkat akurasi yang lebih tinggi daripada algoritma Naïve Beyes dalam segi penyesuaian jurusan dan rekomendasi jurusan dengan tingkat akurasi 93.31% dengan rekomendasi akurasi jurusan sebesar 82,64.

Tabel 2.1 Penelitian Terkait

No	Nama Peneliti dan Tahun	Masalah	Metode	Hasil
1.	Yudhi Andrian, M. Rhifky Wayahdi, 2014	Bagaimana Memprediksi minat siswa yang mendaftar di kampus ABC secara optimal	Bagaimana Memprediksi minat siswa yang mendaftar di kampus ABC secara optimal	ni50 data diuji menghasilkan 40 siswa yang mendaftar dan 10 siswa yang tidak mendaftar pada proses gain tertinggi proses klasifikasi algortima c4.5 dalam atribut lai akurasinya dapat mencapai 83,57%, dan terus meningkat hingga 87,63%
2.	Selvia Lorena Br Ginting, Wendi Zarman, Ida Hamidah, 2014	Tingkat kelulusan sarjana komputer semester 8 sangat kecil	Pembentukan pohon keputusan dengan algoritma C4.5	Data training menentukan tingkat kecocokan kecil atau besar dengan data training 110 akan memiliki

No	Nama Peneliti dan Tahun	Masalah	Metode	Hasil
		yaitu hanya 1 orang yang lulus tepat waktu		tingkat kecocokan lebih kecil daripada data training 70 akurasi 87,45%
3	Liliana Swastina, 2013	bagaimana menentukan jurusan yang tepat pada siswa SMU kelas XII	algoritma C4.5 dan algoritma Naïve Beyes	daripada algoritma Naïve Beyes dalam segi penyesuaian jurusan dan rekomendasi jurusan dengan tingkat akurasi 93.31% dengan rekomendasi akurasi jurusan sebesar 82,64

Dari penjelasan tiga jurnal diatas memiliki kaitan dengan penelitian yaitu jurnal pertama menentukan pohon keputusan untuk memperoleh minat suatu objek, jurnal kedua yaitu dalam algoritma C4.5 membutuhkan data training sebagai perhitungannya, jurnal ketiga yaitu dalam menentukan algoritma yang dipakai membutuhkan suatu tingkat akurasi apakah optimal penggunaan algoritma C4.5 ini dan juga bisa dijadikan sebagai perbandingan antar algoritma lainnya.

2.2 Data Mining

2.2.1 Pengertian Data Mining

Data mining memiliki pengertian suatu proses dengan menggunakan teknik pembelajaran komputer satu atau lebih dimana untuk menganalisis dan

mengekstrak pengetahuan dilakukan secara otomatis yang didalamnya menerapkan metode saintifik atau yang disebut KDD (Knowledge Discovery in Database . Data mining memiliki tujuan yakni data data set yang tersedia dapat diprediksi dengan tipe biner atau nominal. Adapun komponen komponen dari proses data mining yaitu [4]:

1. Kelas

   Definisi kelas ialah suatu bagian label dari hasil klasifikasi yang didalamnya menggunakan variabel tidak bebas.

2. Prediktor

   Prediktor merupakan model yang variabelnya bebas dari atribut data yang telah diklasifikasi.

3. Set Data Pelatihan

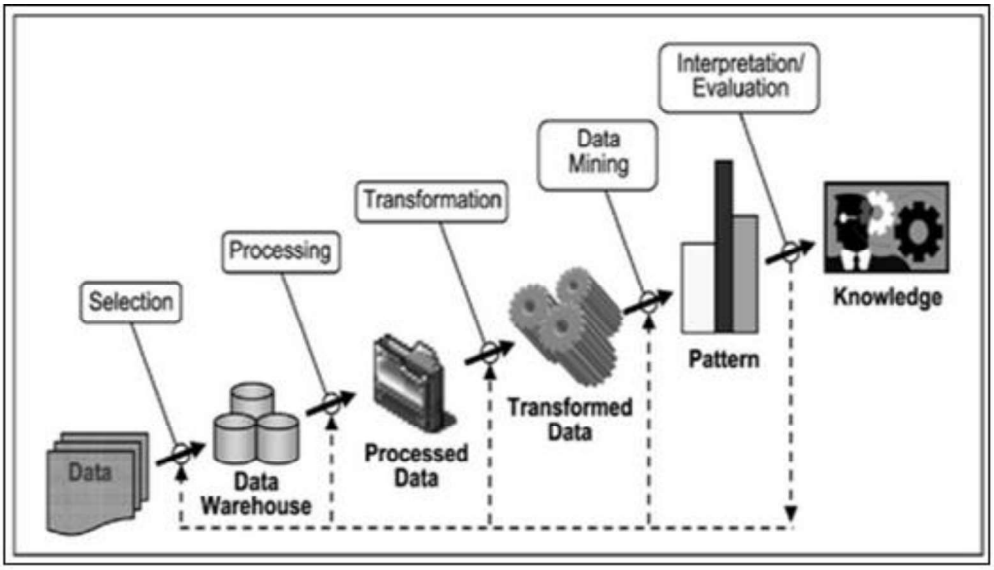
   Merupakan data yang banyak dan lengkap yang didalamnya memiliki kelas dan predictor kemudian dilatih dari model tersebut dikelompokan dalam kelas yang tepat.

4. Set Data Uji

   Set data uji ialah data yang tergolong baru yang nantinya dikelompokan ke dalam model yang berguna mengetahui tingkat akurasi dari modal yang sudah dibuat.

2.2.2 Proses Data Mining

Suatu pembahasan ilmu pasti tidak terlepas dari proses proses yang harus dijalankan, KDD adalah pengetahuan yang masih luas yang masih belum diketahui dan salah satu dalam proses KDD ada data mining, berikut penjelasan KDD khususnya kaitannya dengan data mining [5].



Gambar 2.1 Process Knowledge Discovery in Database

Dari gambar tersebut adapun penjelasan proses KDD antara lain :

1. Data selection

   Kumpulan kumpulan data yang diseleksi sebelum ketahap menggali semua informasi KDD didalamnya.

2. Pre-Processing

   Suatu langkah pembersihan data data yang dianggap tidak penting, data yang tidak sempurna, dan data yang tidak valid agar meningkatkan mutu atau akurasi data.

3. Transformation

   Suatu teknik yang membutuhkan format data yang khusus sebelum diimplementasikan seperti data numerik yang harus dibagi bagi menjadi beberapa interval untuk menentukan kualitas tergantung dari karakteristik teknik teknik data mining tertentu.

4. Data Mining

   Prose mencari pola menarik atau informasi data yang menarik untuk diolah melalui teknik teknik atau metode tertentu.

5. Interpretation (Evaluation)

   Dari proses data mining memerlukan suatu bentuk yang dapat dimengerti oleh pihak yang bersangkutan dan menyelidiki apakah informasi benar atau bertentangan dengan hipotesis.

2.2.3 Pengelompokan Data Mining

Pengelompokkan data mining berdasarkan tugasnya antara lain [5]:

1. Deskripsi

   Menjelaskan semua fakta dan keterangan dari suatu objek.

2. Estimasi

   Model yang dibangun dengan menggunakan record yang didalamnya terdapat nilai dari variabel prediksi yang variable estimasinya kearah numerik.

3. Prediksi

   Suatu hasil data dari objek yang diprediksi untuk menghasilkan informasi dimasa depan.

4. Klasifikasi

   Pengelompokan atribut berdasrakan kategori yang sama.

5. Pengklusteran

   Pengelompokkan record untuk membentuk suatu kelas objek objek yang hampir sama kemiripannya objek lainnya dan tidak mempunyai kemiripan dengan record dalam cluster lainnya.

6. Asosiasi

   Mencari atribute dalam kurun waktu tertentu.

2.3 Klasifikasi Data Mining

Klasifikasi memiliki definisi yaitu proses yang didalamnya mempelajari fungsi memiliki suatu tujuan dimana tiap himpunan dilakukan pemetaan dari himpunan x

ke satu dari label kelas y dari definisi sebelumnya. Model klasifikasi mempunyai 2 jenis antara lain [4]:

1. Pemodelan Deskriptif

   Fungsi dari pemodelan ini ialah memebedakan suatu objek kedalam suatu kelas kelas yang berbeda.

2. Pemodelan Prediktif

   Fungsi dari pemodelan ini ialah memprediksi label kelas record yang masih belum diketahui.

2.4 Algoritma C4.5

Definisi algoritma C4.5 yaitu metode klasifikasi data yang memiliki dua tipe antara lain tipe kategorikal (nominal atau ordinal) dan tipe numerik (interval atau rasio) dengan melakukan pemotongan (pruning) untuk memperoleh gain sehingga membentuk decision tree atau pohon keputusan [6]. Rumus untuk algoritma C4.5 yaitu:

Rumus Untuk Menghitung Gain:

$$Gain\ (S, A) = Entropy\ (S) - \sum_{i=-1}^{n} \left|\frac{si}{s}\right| \times \text{Entropy}(Si) \quad (2.1)$$

Dengan:

S : Himpunan kasus

A : Atribute

N : Jumlah partisi atribute A

$|S_i|$: Jumlah kasus pada partisi ke i

|S| : Jumlah kasus dalam S

Rumus menghitung Entropy:

$$Entropy\ (S) = \sum_{i-1}^{n} -pi^{*}\ log_2 pi \quad (2.2)$$

Dengan:

S : Himpunan kasus

A : Fitur

N : Jumlah partisi S

pi : Proporsi dari S_i terhadap S

2.5 Decision Tree

Pohon keputusan yang digunakan untuk prosedur penalaran yang menghasilkan jawaban dari masalah yang dimasukkan [6]. Elemen-elemen atau node desicion tree, yaitu:

1. Node akar

   Node yang karakteristiknya tidak mempunyai lengan masukkan, dan mempunyai lengan keluaran 0 atau lebih.

2. Node Internal

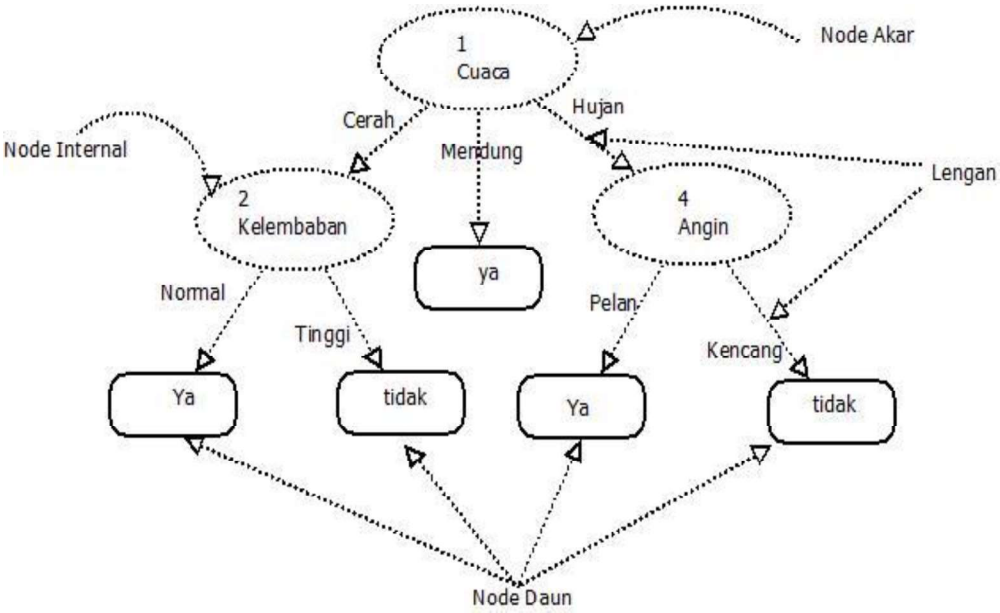
   Node yang mempunyai karakteristik nonterminal atau bukan daun yang memiliki 1 lengan masukan dan memiliki keluaran dua atau lebih.

3. Lengan

   Suatu cabang dari setiap bagian cabang untuk mengungkapkan pernyataan dari nilai hasil pengujian pada node.

4. Node Daun (Terminal)

   Suatu node dengan karakteristik mempunyai satu lengan masukan dan tidak ada lengan keluaran dengan menyatakan label kelas atau keputusan. Berikut contoh gambar decision tree:



Gambar 2.2 Decision Tree

2.6 Confusion Matrix

Confusion matrix merupakan metode pengujian yang digunakan untuk mengukur kinerja suatu model klasifikasi [3]. Pengujian confusion matrix dilakukan dengan menggunakan data testing yang dicocokan dengan hasil dari klasifikasi data training maka akan mendapatkan akurasi, recall, dan Precision. Confision matrix menggunakan table yang berisi benar atau tidak benar dari hasil data testing yang telah dicocokkan dengan klasifikasi data training. Contoh dari tabel confusion matrix dapat dilihat dari tabel 1 confusion matrix.

Tabel 2.2 Confusion Matrix

		PREDICTED CLASS	
		CLASS = 1	CLASS = 0
ACTUAL CLASS	CLASS = 1	F11	F10
	CLASS = 0	F01	F00

Dimana penjelasannya sebagai berikut :

F11 = Jika hasil prediksi positif dan data sebenarnya positif

F10 = Jika hasil prediksi negatif dan sedangkan nilai sebenarnya positif

F01 = Jika hasil prediksi positif dan sedangkan nilai sebenarnya negatif

F00 = Jika hasil prediksi negatif dan data sebenarnya negatif

$$Akurasi = \frac{F11 + F00}{F11 + F10 + F01 + F00} \quad (2.3)$$

Sedangkan precision merupakan ukuran dari akurasi suatu kelas tertentu yang telah diprediksi. Adapun rumus untuk mecari precision adalah sebagai berikut:

$$Precision = \frac{F11}{F01 + F11} \quad (2.4)$$

Sedangkan recall merupakan persentase dari data dengan nilai positif yang nilai prediksinya juga positif. Adapun rumus untuk mencari recall adalah sebagai berikut:

$$Recall = \frac{F11}{F10 + F11} \quad (2.5)$$

2.7 Perpustakaan

Perpustakaan adalah tempat dimana kumpulan buku ilmu pengetahuan yang banyak dan ditata rapi sesuai jenis pengetahuannya yang berguna untuk dibaca, dipelajari dan sebagai bahan rujukan. Sedangkan Minat baca adalah sesuatu kemampuan manusia yang terjadi secara tidak otomatis yang harus didahului dengan kebiasaan-kebiasaan membaca. Minat membaca berupa dari media cetak seperti majalah, buku pelajaran. Koran dan lain lain yang nantinya memiliki daya tarik sendiri setiap judul yang ingin dibacanya dengan berulang kali [7].

2.8 PHP

Menurut Arief (2011) PHP adalah sebuaha bahasa pemrograman berupa bahasa server-side–scripting dan digabung dengan HTM yang digunakan untuk membuat web yang dinamis. Dan PHP tersebut berupa server-side–scripting sehingga perintah-perintah yang ada dalam php tersebut di eksekusi oleh server kemudian hasil tersebut dikirimkan ke browser dan memakai format HTML [8].

Untuk membuat sebuah web masih ada lagi contoh bahasa pemrograman selain PHP karena kita dapat membuat web hanya dengan HTML saja. Dengan demikian web yang dibuat melalui HTML dan css sering disebut web statis yaitu konten dan halaman web tersebut bersifat tetap.

Website yang bersifat dinamis yang dapat dibuat dengan memakai PHP sebuah situw web yang bisa dibilang konten dalam halaman web tersebut dapat menyesuaikan situasi yang ada. Website dinamis ini juga berguna untuk penyimpanan data pada database, membuat halaman dapat diubah-ubah sesuai keinginan inputan dari user, pemrosesan form dan lain lain. Dalam pembuatan web, biasanya kode PHP sering disisipkan pada dokumen HTML. Sehingga PHP memiliki fitur tersebut sering dinamakan sebagai Scripting Language atau bahasa pemrograman script. Beberapa keunggulan PHP daripada bahasa pemrograman komputer yang lainnya antara lain:

1. Server-server pada webiste telah mendukung pada bahasa pemrograman PHP ini.
2. Bagi developer mudah mempelajari bahasa pemrograman php karena pada kode kode bahasa dan script dapat mudah untuk dipahami.
3. PHP merupakan Bahasa Pemrograman yang tidak pernah dan tidak akan pernah melakukan sebuah kompilasi didalam penggunannya.
4. PHP sangat didukung dengan adanya referensi-referensi yang banyak dari berbagai sumber.
5. Bahasa pemrograman PHP dapat dijangkau pada semua sistem operasi seperti UNIX, Linux, Windows, dan lain sebagainya.
6. Bahasa Pemrograman PHP Dapat menjalankan sebuah ataupun beberapa perintah dari suatu sistem.
7. PHP dapat dijalankan dan digunakan secara runtime melalui sebuah konsol.

2.9 MYSQL

MYSQL adalah sebuah database secara multiuser dengan menggunakan bahasa pemrograman. Sedangkan SQL adalah bahasa pemrograman berguna untuk mengakses server database. Dibandingkan dengan yang lain SQL lebih friendly dalam mengakses database daripada dBase atau Clipper karena menggunakan perintah dari pemrograman [9]. MYSQL memiliki kelemahan maupun kelebihan diantaranya sebagai berikut:

1. Kelebihan MySQL

   Adapun kelebihan MySQl dalam penggunaanya dalam database adalah:

   a. Untuk mendapatkan MYSQL sangatlah mudah karena bersifat free atau dapat diunduh secara gratis.
   b. Dalam pengoperasiannya MySQl sangatlah stabil dan tangguh.
   c. Sistem keamanan yang dimiliki MYSQL termasuk cukup baik.
   d. Memiliki dukungan dari semua komunitas dan transaksi.
   e. Support untuk berbagai amcam program untuk menjalankannya.
   f. MySQL memiliki perkembangan yangs angat cepat.

2. Kelemahan MySQL

   Selain itu MYSQL juga memiliki beberapa kelemahan yang patut dibahas antara lain:

   a. Kurang mendukung koneksi bahasa pemrograman seperti Visual basic atau biasa kita kenal dengan sebutan VB, Foxpro, Delphi dan lain-lain sebab koneksi ini menyebabkan field yang dibaca harus sesuai dengan koneksi dari bahasa pemrograman visual tersebut.
   b. Data yang dapat ditangani belum besar dan belum mendukung widowing function.