BAB II
TINJAUAN PUSTAKA DAN LANDASAN TEORI

2.1. Tinjauan Pustaka

(Kamagi, 2014) dalam penelitiannya peneliti menggunakan 3 atribut yaitu jenis kelamin, asal sekolah SMA dan indeks prestasi dari semester 1 sampai 6. Dalam penelitian tersebut digunakan 100 data alumni dari angkatan 2007 sampai 2008 sebagai data training sedangkan dari angkatan 2009 sebagai data testing. Dari data–data tersebut diolah menggunakan algoritma C4.5 sehingga didapatkan hasil prediksi kelulusan terhadap data testing tingkat akurasi sebesar 87.5% yang dipengaruhi oleh atribut indeks prestasi semester 6.

Berdasarkan penelitian ini menggunakan metode Fase CRISP-DM (*Cross Industry Standard Process for Data Mining*). Penelitian ini menggunakan 9 atribut yaitu Nomor Induk Mahasiswa (NIM), Nama, Jenjang, Program Studi, Nama Provinsi, Jenis Kelamin, SKS yang telah ditempuh, IPK, dan tahun kelulusan. Dalam penelitian ini peneliti menggunakan metode klasifikasi *naive bayes*. Dataset yang digunakan adalah data mahasiswa jenjang D-3 dan S1 pada tahun 2009 dan data jumlah kelulusan tahun 2011 hingga 2013. Pada penelitian ini terdapat pembahasan pemilihan atribut dari atribut dataset yang ada. Untuk melakukan pengolahan data peneliti menggunakan Rapidminer sebagai alat pengolahan data menggunakan algoritma *naive bayes* dan didapatkan hasil dengan tingkat akurasi 82.08%.(Nugroho, 2014)

Jananto (2013) Dalam studi kasus yang dilakukan di Program Studi Sistem Informasi, Universitas Stikubank. Penelitian ini menggunakan metode klasifikasi *naive bayes* dengan membagi tahap masa evaluasi nilai semester 4 sebagai tahap 1. Penelitian ini menggunakan metode klasifikasi *naive bayes* dengan membagi tahap masa evaluasi nilai semester 4 sebagai tahap 1. Penelitian ini menggunakan 5 atribut antara lain kota lahir, jenis sekolah, kota sekolah, jenis kelamin, dan data akademik sampai semester 4. Dilakukan *record* data sebanyak 266 *record* diperoleh klasifikasi dengan tingkat kesalahan klasifikasi sebanyak 20 *record* dari total sebanyak 66 *record* data testing atau 34%. Dalam penelitian ini peneliti melakukan 3 kali tahap uji coba dengan pengujian yang berbeda – beda.

(Nugroho, 2015) Penelitian ini melakukan analisis data mahasiswa sebagai sumber keputusan untuk menentukan calon mahasiswa baru berdasarkan kriteria tertentu. Dalam penelitian ini pada tahap klastering menggunakan algoritma *K-Means* yang kemudian diklasifikasikan menggunakan metode *Decision Tree*. Data yang digunakan dalam penelitian ini yaitu data alumni sejumlah 223 data mahasiswa dan 209 data mahasiswa yang masih aktif. Dua tahap klasifikasi yang digunakan untuk menentukan klasifikasi mahasiswa berdasarkan masa studi dan pola predikat kelulusan. Penelitian ini menggunakan rapidminer sebagai alat untuk pengolahan data. Hasil dari prediksi ini didapat beberapa kriteria dimana variabel yang paling berpengaruh terhadap predikat kelulusan mahasiswa adalah partisipasi mahasiswa menjadi asisten.

Dalam penelitian ini melakukan analisis untuk melihat faktor – faktor penyebab terjadinya kecelakaan kerja menggunakan algoritma C4.5 yang kemudian hasilnya digunakan sebagai panduan untuk menghindari resiko kecelakaan *(zero accident)*, agar kualitas dan kuantitas pekerjaan menjadi baik dan memenuhi target. Faktor yang digunakan dalam proses pengambilan keputusan yang diidentifikasi adalah lingkungan tempat kerja, alat pelindung diri, pekerja dan cara kerja, material, rambu – rambu keselamatan dan variabel yang digunakan yaitu Baik dan Tidak Baik. Peneliti melakukan proses perhitungan menggunakan algoritma C4.5 secara manual. Dari hasil pembahasan dan pengujian yang dilakukan dapat ditarik kesimpulan bahwa informasi atau pengetahuan tentang faktor – faktor penyebab kecelakaan kerja kontruksi yang terjadi pada proyek PT. Arupadhatu Adisesanti adalah pekerjaan dan cara kerja, lingkungan tempat kerja, dan alat pelindung diri. Metode algoritma C4.5 atau pohon keputusan lebih efektif dan fleksibel jika digunakan pada proses pengklasifikasian. Penelitian ini telah diuji menggunakan *software* Data Mining *WEKA* Gui Chooser dan memiliki hasil pengklasifikasian sama.(Erlin, 2017)

(Handhayani, 2017) Pada penelitian ini dilakukan analisis data menggunakan nilai mahasiswa untuk memprediksi lama waktu studi mahasiswa dan kinerja mahasiswa. Lama studi mahasiswa sangat penting untuk akademik dan institusi untuk membantu mahasiswa dalam merancang atau mengatur studinya. Dataset yang digunakan antara lain jenis kelamin, daerah asal, pekerjaan orang tua, IPK, nilai yang dicapai, dan nilai dari perencanaan serta kontrolnya. Penelitian ini menggunakan 3 klaster yaitu mahasiswa yang

kurang, mahasiswa rata – rata, dan mahasiswa pintar. Selain itu penelitian ini menjadikan 2 kelompok atau kategori lama study mahasiswa dengan masa studi 3,5 sampai 4 tahun sebagai kelompok tepat waktu dan masa studi 5 sampai 7 tahun disebut sebagai kelompok terlambat. Dalam penelitian ini menggunakan *mutual information* (MI) dan menggunakan fitur – fitur yang memiliki relasi kuat dengan masa studi mahasiswa. Penelitian ini melakukan 2 percobaan yang pertama menggunakan 2 kelas target dan yang kedua menggunakan 3 kelas target. Kelas target dikelompokkan dalam 4 tabel berdasarkan lama studinya. Dataset menggunakan 240 sampel dan 25 subjek atau fitur. Dihasilkan 8 tabel yang berelasi dengan beberapa kelas. Hasilnya 69,17% tergolong dalam kelompok tepat waktu dan 30,83% tergolong dalam kelompok terlambat. Percobaan ini diulang sebanyak 50 kali dengan menggunakan 70% data training dan 30% data testing. Teknik ini digunakan ke semua algoritma. Hasil dari proses pemilihan fitur ada 12 subjek yang melalui 2 bagian percobaan. Bagian pertama kinerja SVM menghasilkan 83,64% akurasi, setelah pemilihan fitur akurasi naik mencapai 85,64%. Bagian kedua menggunakan 3 kelas target dengan keakuratan SVM sekitar 77% dan 80% yang sebelumnya telah dipilih fitur masing–masingnya. Seleksi fitur MI berhasil diimplementasikan memilih subjek yang memiliki hubungan dengan kelas target.

Pada penelitian dijelaskan bahwa pemantauan mahasiswa masuk, perkembangan mahasiswa, prestasi mahasiswa, rasio kelulusan dari jumlah mahasiswa yang lulus, dan kompetensi dari lulusan harus mendapatkan perhatian serius untuk menerima atau menghargai kepercayaan dan kebutuhan alumni. Metode yang digunakan peneliti yaitu klasifikasi menggunakan algoritma *Naive bayes*. Teknik merupakan salah satu program utama di IBI Darmajaya yang tergolong dalam pilihan utama favorit di tahun 2008–2011 dengan rata–rata 250 mahasiswa setiap tahun. Pada tahun 2012–2014 mengalami degradasi dilihat dari kuantitas minat mahasiswa. Selain itu tingkat kelulusan rata–rata mengalami penurunan. Atribut yang digunakan antara lain jenis kelamin, kota asal, tipe sekolah, lokasi sekolah, ekonomi, IPK, dan kategori. Dalam penelitian ini untuk memonitor perkembangan studi mahasiswa dan mengantisipasi mahasiswa yang kinerjanya kurang. Data sampel yang digunakan adalah data mahasiswa yang telah lulus sebagai data training. Data training tersebut merupakan data angkatan 2011–2012. Peneliti membagi kategori data menjadi 3

yaitu cepat, tepat waktu, dan terlambat. Dari 191 data mahasiswa 50 *record* yang digunakan sebagai data training. Setelah diproses, didapat hasil 21 mahasiswa masuk dalam kategori cepat, 6 mahasiswa dalam kategori tepat waktu dan 23 mahasiswa dalam kategori terlambat. Proses yang dilakukan dibagi menjadi 2 yaitu training dan testing. Algoritma yang digunakan dalam testing adalah algoritma *naive bayes*. Setelah proses testing didapat hasil dari 50 data training yang diimplementasikan dari setiap kategori didapat total 42% cepat, 12% tepat waktu, dan 46% terlambat. Selanjutnya diambil 20 data testing didapat hasil 20% cepat, 35% tepat waktu, dan 45% terlambat. Dari dua hasil testing tersebut ditarik kesimpulan bahwa point paling tinggi terdapat pada kategori terlambat. Dari setiap kategori yang telah dibandingkan didapat atribut jenis kelamin dan asal sekolah yang paling mempengaruhi (Artaye, 2015).

(Setiawan, 2015) Penelitian ini merupakan penelitian yang dilakukan untuk menjaga kualitas pendidikan dan hasilnya melewati proses evaluasi. Membuat prediksi kelulusan siswa dan menentukan faktor-faktor yang menghambat sebagai masukan untuk universitas. Untuk membuat segmentasi kinerja mahasiswa dan memprediksi kelulusannya dapat dilakukan dengan klasifikasi diagram kuadran dari 2 parameter yaitu IPK dan IPS hingga semester 4. Algoritma C5, C&R *Tree*, CHAID, dan regresi logis diuji untuk melihat model yang paling tepat. Hasil akhir yang diharapkan yaitu dapat mengkategorikan mahasiswa berdasarkan kinerja dan memprediksi kelulusannya agar penelitian ini dapat menjadi bahan pertimbangan universitas dalam mengambil tindakan untuk meningkatkan IPK dan tingkat kelulusan. Penelitian ini menggunakan metodologi standar data mining yaitu CRISP-DM. Setelah melakukan persiapan data dan pemodelan didapatkan model yang paling baik dari model yang ada yaitu IBM SPSS modeler menggunakan *auto-classifier node* yang merupakan teknik yang kuat untuk mengestimasikan dan menggabungkan nilai – nilai daripada pemodelan yang lainnya dan memilih C5 untuk memprediksi keberhasilan kurikulum tepat waktu. Keakuratan prediksi yang dihasilkan dari penyelesaian data kurikulum dari 87% menjadi 88,034%. Hal ini menunjukkan bahwa model yang digunakan berhasil. Namun hasil dari IBM SPSS modeler ini akan menghasilkan akurasi yang berbeda apabila diterapkan pada kelompok data yang berbeda, misalnya untuk universitas lainnya, karena parameter hasil nilainya akan berbeda. Metode model yang diterapkan telah diuji,

data set yang berbeda harusnya menjadi pemodelan ulang, untuk membentuk sesuai model. Hasil prediksi dan rekomendasi dapat digunakan oleh konselor dalam memotivasi siswa untuk mengintensifkan upaya untuk mencapai keberhasilan perkuliahan. Untuk akurasi prediksi, informasi eksternal seperti kegiatan akademik dan latar belakang keluarga dimasukkan sebagai bagian dari data demografi mahasiswa.

Berdasarkan penelitian yang dilakukan untuk memperoleh mahasiwa yang akan mewakili lomba pada kompetisi *event* cyberjawara yang diselenggarakan oleh kementrian Komunikasi dan Informatika RI. Kemampuan yang dibutuhkan untuk mengikuti kompetisi tersebut yaitu kemampuan untuk melakukan analisis untuk memecahkan permasalahan terkait logika pemrograman agar bisa meraih juara yang diharapkan. Variabel–variabel yang dijadikan acuan untuk memenuhi kriteria yang baik antara lain adalah nilai matakuliah Algoritma dan Pemrograman, Fisika Dasar, Kalkulus dan IPK. Metode yang digunakan yaitu *K-Means* untuk menentukan pengelompokkan mahasiswa dengan kriteria tersebut. *Software* yang digunakan dalam penelitian ini adalah *WEKA* yang tujuannya untuk membandingkan hasil dengan perhitungan secara teoritis dengan hasil yang didapatkan dengan proses di mesin *WEKA*. Jumlah instance yang digunakan adalah 124 dari 5 atribut yang digunakan. Dari data yang diuji didapat 4 kelompok atau 4 *cluster* dengan perolehan 7% di cluster 0, 23% di cluster 1, 32% di cluster 2 dan 38% di cluster 3. Maka cluster 1 dengan perolehan IPK tertinggi bisa digunakan untuk memilih 5 mahasiswa untuk bisa mewakili lomba. (Asroni, 2016).

(Asril, 2015) Penelitian ini bertujuan untuk menganalisa dan mengolah data mahasiswa yang telah lulus untuk mendapatkan informasi yang penting dan bermanfaat yang mempermudah pihak pemasaran universitas dalam melakukan promosi dan mencari calon mahasiswa baru diberbagai kota di Indonesia. Atribut yang digunakan yaitu nama mahasiswa, jurusan yang diambil, sekolah asal, kota asal mahasiswa, dan IPK. Metode data mining yang digunakan adalah klastering menggunakan algoritma *K-Means*. Pada penelitian ini ada 3 tahap pokok yang dilakukan peneliti yaitu pra proses, proses, dan presentasi. Pra proses meliputi pemilihan dan penyeleksian data dari data mentah agar bisa digunakan untuk proses mining. Sedangkan proses merupakan tahap penggunaan algoritma *K-Means* pada data. Jumlah klaster yang digunakan adalah 4 klaster. Kemudian presentasi

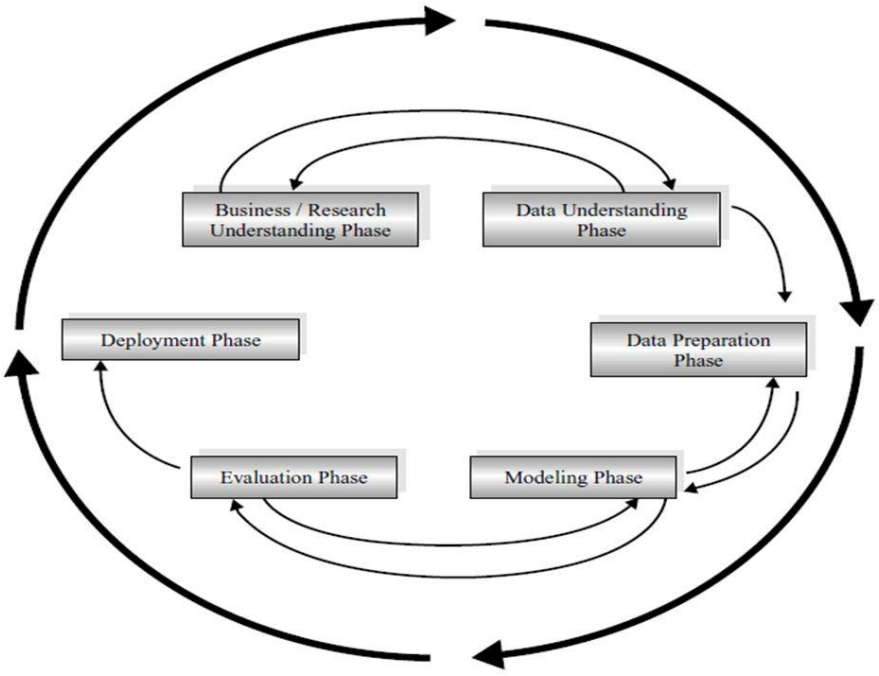
atau hasil. Dari data hasil pengolahan data dengan algoritma dapat ditentukan beberapa strategi promosi bagi calon mahasiswa baru untuk setiap wilayah berdasarkan cluster yang terbentuk dengan mengirim tim promosi berdasarkan potensi akademik mahasiswa dengan melihat IPK pada setiap cluster.

2.2. Landasan Teori

2.2.1. Data Mining

Data mining adalah proses untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dengan berbagai *database* besar menggunakan teknik statistik, matematika, kecerdasan buatan dan machine learning (Turban, 2005)

Menurut (Kusrini, 2009) Pada tahun 1996 beberapa analis industri mengembangkan Cross-Industry Standard Process for Data Mining (CRISP-DM). CRISP DM menyediakan standar proses data mining sebagai stategi dalam memecahkan masalah secara umum dari bisnis atau penelitian. Dalam CRISP-DM proyek data mining memiliki siklus hidup yang terbagi dalam enam fase (gambar). Keseluruhan fase yang berurutan tersebut bersifat adaptif.



Gambar 2.1 Proses data mining menurut CRISP-DM

Enam fase CRISP–DM (Larose,2005):

- Fase Pemahaman Bisnis (*Business Understanding Phase*)

  Fase ini merupakan fase penentuan tujuan dan kebutuhan secara detail dan keseluruhan. Fase ini juga fase dimana tujuan dan batasan yang menjadi formula dari permasalahan data mining diterjemahkan dan fase dalam menyiapkan stategi awal untuk mencapai tujuan.
- Fase Pemahaman Data (*Data Understanding Phase*)

  Fase ini adalah fase untuk mengumpulkan data dan menggunakan analisis penyelidikan data untuk pengenalan lebih lanjut dan pencarian pengetahuan awal. Fase ini merupakan fase untuk mengevaluasi kualitas data dan memilih group data yang mengandung pola dari permasalahan yang diinginkan.
- Fase Pengolahan Data (*Data Preparation Phase*)

  Fase ini merupakan fase persiapan dari data awal untuk keseluran fase berikutnya dan untuk perangkat pemodelan. Memilih kasus dan mengubah beberapa *variabel* yang dibutuhkan untuk dianalisis dan sesuai dengan analisis yang akan dilakukan.
- Fase Pemodelan (*Modelling Phase*)

  Ini merupakan fase memilih dan pengaplikasikan teknik pemodelan yang sesuai, kalibrasi aturan model untuk mengoptimalkan hasil. Jika diperlukan, proses dapat kembali ke fase pengolahan data untuk menjadikan data ke dalam bentuk yang sesuai dengan spesifikasi kebutuhan teknik data mining tertentu.
- Fase Evaluasi (*Evaluation Phase*)

  Fase untuk mengevaluasi satu atau lebih model yang digunakan dalam fase pemodelan untuk mendapatkan kualitas dan efektivitas sebelum disebarkan untuk digunakan. Menetapkan model yang memenuhi tujuan fase awal, menentukan permasalahan penting yang tidak tertangani dengan baik, dan mengambil keputusan berkaitan dengan penggunaan hasil dari data mining.
- Fase Penyebaran (*Deployment Phase*)

Fase penggunaan model yang dihasilkan. Terbentuknya model tidak menandakan telah terselesainya proyek, penerapan data mining secara parallel pada departemen lain dan pembuatan laporan juga dibutuhkan untuk mendukung penyempurnaan proyek.

Data mining dibagi menjadi beberapa kelompok berdasarkan tugas yang dapat dilakukan, yaitu (Larose, 2005):

- Deskripsi

  Terkadang peneliti atau analis secara sederhana ingin mencoba mencari cara untuk menggambarkan pola dan kecenderungan yang terdapat dalam data. Deskripsi dari pola dan kecenderungan sering memberikan kemungkinan penjelasan untuk suatu pola atau kecenderungan.
- Estimasi

  Estimasi hampir sama dengan klasifikasi kecuali variabel target. Estimasi lebih kearah numerik daripada kearah kategori. Model yang dibangun menggunakan *record* lengkap yang menyediakan nilai dari variabel target sebagai nilai prediksi. Selanjutnya, pada peninjauan berikutnya estimasi nilai dari variabel target dibuat berdasarkan nilai variabel prediksi.
- Prediksi

  Prediksi hampir sama dengan klasifikasi dan estimasi kecuali dalam prediksi nilai dari hasil aka nada di masa mendatang. Beberapa metode dan teknik yang digunakan dalam klasifikasi dan estimasi dapat pula digunakan (untuk keadaan yang tepat) untuk prediksi.
- Klasifikasi

  Pada klasifikasi terdapat target variabel kategori yang merupakan teknik mengklasifikasi data. Pada klasifikasi diharuskan ada variabel dependen sedangkan pada klastering variabel dependen tidak ada.
- Pengklusteran

  Pengklusteran merupakan pengelompokkan *record*, pengamatan, atau memperhatikan sehingga membentuk kelas objek–objek yang memiliki kemiripan. Kluster adalah kumpulan *record* yang memiliki kemiripan satu

dengan lainnya dan memiliki ketidakmiripan dengan *record–record* dalam kluster lain. Perbedaan pengklusteran dengan klasifikasi yaitu tidak adanya variabel target dalam pengklusteran. Pengklusteran tidak melakukan klasifikasi, mengestimasi atau memprediksi nilai dari *variabel* target. Tetapi pengklusteran melakukan pembagian atau pengelompokkan terhadap keseluruhan data yang memiliki kemiripan (homogen), yang mana kemiripan *record* dalam suatu kelompok akan bernilai maksimal sedangkan kemiripan dengan *record* dalam kelompok lain akan bernilai minimal.

- Asosiasi

  Asosiasi dalam data mining adalah menemukan atribut yang muncul dalam satu waktu atau dalam dunia bisnis lebih umum disebut analisis keranjang belanja.

2.2.2. Algoritma C4.5

(Kamagi, 2014) Secara umum untuk membangun pohon keputusan Algoritma C4.5 adalah sebagai berikut :

a. Pilih atribut sebagai akar
b. Buat cabang untuk masing–masing nilai
c. Bagi kasus dalam cabang
d. Ulangi proses masing–masing cabang hingga semua kasus memiliki kelas yang sama.

Pemilihan atribut sebagai akar didasarkan pada nilai gain tertinggi dari atribut–atribut yang ada. Untuk menghitung gain digunakan rumus seperti berikut:

$$\text{Gain}(S,A) = \text{Entrophy}(s) - \sum_{i=1}^{n} - \frac{|si|}{|s|} * \text{Entrophy}(si)$$

Persamaan 2.1

Keterangan:

S : Himpunan kasus

A : Atribut

n : Jumlah partisi atribut A

|Si| : Jumlah kasus pada partisi ke i

|S| : Jumlah kasus dalam S

Untuk mendapatkan nilai Gain hal yang harus dilakukan sebelumnya yaitu mencari nilai Entropi. Entropi digunakan untuk menentukan seberapa informatif atribut masukan untuk menghasilkan sebuah atribut. Rumus dasar dari entropi adalah sebagai berikut:

$$\text{Entrophy (s)} = \sum_{i=1}^{n} - \text{pi} * \log_2 \text{pi}$$

Persamaan 2.2

Keterangan:

S : Himpunan kasus

n : Jumlah partisi S

pi : Proporsi dari Si terhadap S

2.2.3. Algoritma *K-Means*

Algoritma *K-Means* merupakan algoritma klasterisasi dimana data dikelompokkan berdasarkan titik pusat klaster (*centroid*) terdekat dengan data. Tujuannya adalah mengelompokkan data dengan memaksimalkan kemiripan data salam suatu klaster dan meminimalkan kemiripan data antar klaster. Ukuran kemiripan yang digunakan dlam klaster yaitu fungsi jarak. Sehingga kemiripan data yang maksimal didapatkan dari jarak terpendek antara data terhadap titik *centroid*. Tahapan awal proses klasterisasi menggunakan algoritma *K-Means* adalah menentukan titik awal *centroid* yang pada umumnya dibangkitkan secara acak sesuai dengan jumlah klaster yang ditentukan diawal. Setelah *centroid* terbentuk kemudian dihitung jarak tiap datanya. Perhitungan jarak *Euclidean* adalah sebagai berikut (Asroni, 2016):

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Persamaan 2.3

Keterangan :

d : jarak

x : data

y : *centroid*

i : banyaknya data

n : jumlah data