BAB II
TINJAUAN PUSTAKA DAN LANDASAN TEORI

2.1 Tinjauan Pustaka

Salah satu metode penambangan data adalah klasifikasi. Metode klasifikasi memiliki beberapa algoritma dan setiap algoritma klasifikasi pada penambangan data memiliki kelebihan dan kekurangan, sehingga penelitian juga dilakukan untuk menganalisa perbandingan diantara algoritma tersebut. Pengukuran kinerja algoritma data mining dapat dilakukan berdasarkan kriteria antara lain keakuratan, kesempurnaan, konsistensi, kecepatan, dapat dipercaya dan interpretabilitas (Han et al., 2012).

Penambangan data sudah banyak diimplementasikan di beberapa bidang seperti bidang pendidikan, kesehatan, bisnis, jaringan internet, dll. Penelitian yang dilakukan pada jaringan internet untuk deteksi penyusupan atau serangan (*intrusion detection*) dengan menggunakan metode SVM dan penggabungan SVM dengan C5.0, kemudian menganalisa perbandingan kinerja keduanya. Penggabungan ini dilakukan untuk meningkatkan akurasi. Hasil penelitian secara empiris adalah penggabungan SVM dan C5.0 mampu meningkatkan kinerja untuk semua kelas (kelas serangan dan normal) dengan akurasi 100% untuk beberapa jenis serangan sehingga dapat meningkatkan sistem keamanan jaringan (Golmah, 2014).

Dalam bidang pendidikan istilah penambangan data disebut dengan *Educational Data Mining (EDM)*. Algoritma yang digunakan adalah pohon keputusan ID3 dan C4.5 untuk prediksi kinerja siswa. Data siswa pada tahun

pertama dianalisa, kemudian diperoleh informasi yang digunakan untuk prediksi kinerja siswa. Data siswa yang digunakan terdiri dari nama, jenis kelamin, *application* ID, nilai ujian kelas X & XII, nilai saat ujian masuk, kategori dan tipe atau jenis penerimaan. Terdapat dua fase dalam klasifikasi yaitu fase pembelajaran dan fase klasifikasi. Data uji diklasifikasikan kedalam dua kelas yaitu "lulus" atau "gagal". Tingkat akurasi diperoleh dengan membandingkan hasil prediksi dengan hasil sebenarnya yaitu 75,145% untuk kedua algoritma ID3 dan C4.5 (Adhatrao et al., 2013).

Kemudian klasifikasi untuk prediksi tingkah laku mahasiswa, kinerja dalam mengerjakan soal ujian, dan lain-lain. Prediksi ini akan membantu pengajar untuk mengidentifikasi kelemahan mahasiswa dan membantu untuk meraih nilai yang lebih baik. Algoritma yang digunakan adalah pohon keputusan ID3 dan C4.5, kemudian membandingkan keduanya. Hasil penelitian menunjukkan bahwa algoritma C4.5 lebih akurat dari pada algoritma ID3 (Kumar & M.N, 2011).

Penelitian lain dibidang kesehatan dilakukan yaitu sistem pakar untuk mendiagnosa penyakit hewan ternak pada sapi dengan metode *bayesian network*. Sistem pakar ini dirancang untuk dapat menirukan keahlian seorang pakar memecahkan masalah yang ada. Tujuan penelitian ini adalah mampu mendeteksi sedini mungkin penyakit pada sapi berdasarkan gejala penyakit yang diberikan sehingga dapat dilakukan penanganan secara cepat. Sistem ini bekerja dengan cara pengguna memilih gejala yang disediakan sistem kemudian gejala tersebut akan disesuaikan dengan rule yang ada sehingga

pengguna memperoleh hasil diagnosa berupa gejala, tipe penyakit, solusi penanganannya dan nilai probabilitasnya (Tinaliah, 2015).

Untuk melakukan perbaikan kinerja dapat menggabungkan beberapa algoritma berdasarkan kelebihan masing-masing algoritma. Penelitian tahun 2006 dilakukan yaitu kombinasi algoritma Pohon Keputusan dan *Naive Bayes* yang diimplementasikan pada 12 data set. Kombinasi ini dinamakan *Self-adaptif* NBTree. Ukuran pada algoritma Bayes digunakan untuk membangun pohon keputusan dan dapat langsung mengatasi atribut yang kontinu dan otomatis menemukan batas yang tepat untuk proses diskritisasi dan menemukan nilai intervalnya (Wang et al., 2006). Untuk kasus yang lebih spesifik, kombinasi *Decision Tree* dan *Naive Bayes* digunakan untuk mendeteksi serangan pada jaringan. Algoritma kombinasi ini untuk mengatasi kesulitan seperti atribut yang kontinu, nilai atribut yang hilang dan terdapat *noise* (derau) dalam proses pelatihan. Hasil pengujian menunjukkan bahwa kombinasi algoritma *decision tree* dan *naive bayes* mencapai rata-rata deteksi yang tinggi dan secara signifikan mengurangi FP (*False Positives*) untuk tipe gangguan yang berbeda-beda (Farid et al., 2010).

Algoritma klasifikasi yang juga mengkombinasi beberapa algoritma dilakukan pada tahun 2004 yaitu kombinasi algoritma *bayesian network* dan *k-nearest neighbors* untuk analisis data yaitu memprediksi kelas penyakit kanker kedalam tiga data set DNA microarray yaitu Colon, Leukimia dan NCI-60 (Sierra et al., 2004). Selain itu penelitian pada tahun 2013 yaitu kombinasi *naïve bayes classifier* dan *k-nearest neighbor* untuk memprediksi

posisi profitabilitas lembaga keuangan di Negara Bangladesh (Ferdousy et al., 2013).

Klasifikasi data dengan mengkombinasi algoritma *k-nearest neighbor* dan *naive bayes* dilakukan pada tiga data set yang diperoleh dari UCI *machine learning repository* yaitu data set *Nursery*, *Car Evaluation* dan *Balance Scale*. Tujuan dilakukannya kombinasi algoritma ini adalah untuk meningkatkan kinerja proses klasifikasi yaitu mengatasi kelemahan metode KNN dengan waktu lebih cepat dan kelemahan *naïve bayes* dengan persentase akurasi sama atau lebih tinggi (Sari, 2015).

Berdasarkan uraian diatas dari beberapa penelitian yang telah dilakukan, penulis bermaksud untuk melakukan penelitian penambangan data dalam bidang pendidikan dengan menggunakan algoritma klasifikasi yaitu kombinasi algoritma *Bayesian Network* dan *K-Nearest Neighbors*.

Tabel 2.1 Perbandingan Penelitian

Jenis	B.Sierra, dkk (2004)	Farid, dkk. (2010)	Ferdousy, dkk (2013)	Sari (2015)	Windarti (2015) *
Metode	Kombinasi *Bayesian Networks* dan *K-Nearest Neighbors*.	Kombinasi *Decision Tree* dan *Naive Bayes.*	Kombinasi *Naïve Bayes Classifier* and *K-Nearest Neighbor*.	Kombinasi *K-Nearest Neighbors* dan *Naive Bayes.*	Kombinasi algoritma *Bayesian Network* dan *K-Nearest Neighbors*.
Objek Penelitian	Penyakit kanker	Sistem Jaringan	Data set Bank Central	3 data set dari UCI *machine learning repository*	Data alumni mahasiswa UNWIDHA jurusan eksak dan non eksak
Masalah	Bagaimana memprediksi kelas penyakit kanker kedalam tiga dataset DNA microarray dari dimensi data yang besar yaitu dataset Colon, Leukimia dan NCI-60.	Bagaimana mendeteksi gangguan atau serangan pada jaringan dengan menggunakan kombinasi algoritma *Decision Tree* dan *Naive Bayes*. Kemudian bagaimana mengatasi kesulitan yang sering terjadi seperti atribut kontinu, nilai	Bagaimana memprediksi posisi profitabilitas beberapa lembaga keuangan di Bangladesh menggunakan data yang disediakan oleh Bank Central.	Bagaimana mengkombinasikan metode KNN dan *Naïve Bayes* untuk klasifikasi data.	Bagaimana mengkombinasi algoritma *Bayesian Network* dan *K-Nearest Neighbors* berdasarkan kelebihan dan kekurangan. Kemudian bagaimana menganalisa dan membandingkan kinerja kombinasi algoritma dengan *Bayesian Network*,

		atribut yang hilang dan adanya derau (*noise*).			dan *K-Nearest Neighbors* dalam melakukan proses prediksi.
Bahasa Permrograman	-	-	-	Visual C #	Visual C #
Algoritma	1. Cari tetangga terdekat pada kasus baru dalam data latih, menggunakan K-NN, kemudian perluas kasus Ki dengan *Bayesian Network*. 2. Klasifikasikan ke kelas yang memiliki probabilitas posterior terbesar.	1. Klasifikasi data *training* D menggunakan *prior* dan *conditional probbabilities*. 2. Perbaharui nilai kelas pada D. 3. Pilih atribut terbaik A_i dari D dengan nilai information gain maksimum. 4. Bagi data D kedalam sub-datasets. Hitung probabilitas untuk	1. Hitung jarak d(x',x) pada dataset D menggunakan atribut numerik, simpan data yang memiliki nilai jarak terdekat. 2. Membuat model dengan algoritma *NB* menggunakan atribut kategori.	1. Mencari nilai probabilitas dengan Naïve Bayes. 2. Mencari jarak terdekat pada data hasil langkah diatas dengan K-Nearest Neighbors.	1. Hitung posterior probabilitas dengan BN menggunakan atribut kategori. 2. Cari rata-rata nilai probabilitas setiap kelas keluaran, simpan data yang memiliki probabilitas > α (α adalah nilai masukan dari pengguna yaitu 0-1). 3. Hitung jarak antara data uji dengan setiap atribut numerik data latih

		setiap sub-dataset. 5. Pilih atribut terbaik A_i dengan nilai gain maksimum.			menggunakan KNN. 4. Cari jarak terdekat
Tools	-	WEKA untuk mengukur keakuratan kinerja algoritma.	-	-	Ms. Visual Studio 2010 dan Ms. Sql Server 2008.
Tujuan	Mampu memprediksi kelas penyakit kanker ke dalam tiga dataset microarray yaitu Colon, Leukimia dan NCI-60.	Dapat mendeteksi gangguan pada jaringan dengan menggunakan kombinasi algoritma tersebut. Selain itu mampu mengatasi kesulitan yang sering terjadi seperti atribut kontinu, nilai yang hilang dan adanya derau (*noise*).	Mampu meprediksi posisi profitabilitas lembaga keuangan yang ada di Bangladesh melalui data set Bank Central.	Untuk meningkatkan kecepatan waktu dan persentase akurasi dari metode penambangan data KNN dan Naïve Bayes dengan metode kombinasi KNN-Naïve Bayes.	Mampu menggabungkan algoritma *Bayesian Network* dan *K-Nearest Neighbors* dan mampu menganalisa dan membandingkan kinerja penggabungan algoritma dengan *Bayesian Network,* dan *K-Nearest Neighbors*.

Hasil	Menghasilkan kinerja lebih dari yang diharapkan dalam mengklasifikasi data dan hasilnya lebih baik dari algoritma *K-Nearest Neighbors*.	Mencapai rata-rata deteksi yang tinggi dan significant mengurangi FP (*False Positives*) untuk tipe gangguan yang berbeda-beda.	Menghasilkan kinerja yang lebih baik dibanding algoritma klasifikasi lainnya seperti Boost FSNB, DiscNB, WrapperNB, TAN, LBR, BP, SMO, Adaboost C4.5, Bagging C4.5, BaggingNB and AdaboostNB dengan tingkat akurasi 89.58% dengan nilai K=21.	Mengatasi kelemahan pada metode KNN dengan proses waktu lebih cepat dan kelemahan pada *Naive Bayes* dengan persentase akurasi sama dengan atau lebih tinggi.	Harapan tingkat akurasi yang akan dihasilkan untuk prediksi masa studi mahasiswa mencapai minimal 85%.

2.2 Landasan Teori

2.2.1 *Data Mining* (Penambangan Data)

Data mining merupakan proses untuk menemukan pola yang menarik dan pengetahuan dari sejumlah besar data (Han et al., 2012). Sumber data dapat berupa basis data, data warehouse, web, gudang informasi lain dan data yang berasal dari sistem yang dinamis. Selain itu penambangan data juga merupakan proses untuk mendapatkan informasi yang berguna dari gudang basis data yang besar (Tan et al., 2006).

Terdapat beberapa tahapan dalam proses penambangan data antara lain :

1. Pembersihan Data

   Proses untuk mengisi nilai yang hilang, menghilangkan *noise*/derau ketika mengidentifikasi outlier dan memperbaiki data yang tidak konsisten.

2. Integrasi Data

   Penambangan data sering membutuhkan integrasi data. Integrasi data merupakan penggabungan data dari beberapa atau banyak basis data. Integrasi yang dilakukan dengan hati-hati dapat mengurangi dan menghindari terjadinya redundansi dan ketidakkonsistenan dalam menghasilkan data set. Hal ini dapat meningkatkan akurasi dan kecepatan dalam proses penambangan data.

3. Seleksi Data

   Seleksi data dilakukan dengan menganalisa data yang relevan yang akan diambil dari basis data.

4. Transformasi Data

   Dalam transformasi data, data akan diubah atau digabung kedalam bentuk yang sesuai sehingga proses penambangan data menjadi lebih efisien dan pola yang ditemukan akan lebih mudah untuk dimengerti.

5. Proses *mining*

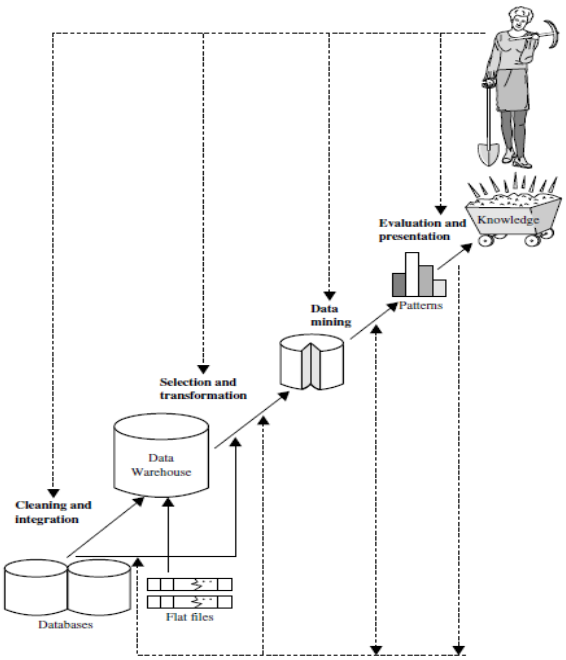
   Proses *mining* merupakan proses yang utama dimana digunakan metode-metode yang berguna untuk mengekstrak pola data.

6. Evaluasi pola

   Digunakan untuk mengidentifikasi pola yang benar-benar menarik yang mewakili pengetahuan berdasarkan pada "*interestingness measures".*

7. Presentasi pengetahuan

   Tahap dimana gambaran teknik visualisasi dan pengetahuan digunakan untuk memberikan pengetahuan yang telah ditambang kepada pengguna.



Gambar 2.1 Tahapan dalam Penambangan Data ((Han et al., 2012)

2.2.2 Klasifikasi

Penambangan data dibagi menjadi enam kelompok yaitu model deskripsi, estimasi, prediksi, klasifikasi, klusterisasi dan asosiasi (Larose, 2006). Klasifikasi merupakan proses untuk menemukan sebuah model atau fungsi untuk menjelaskan dan membedakan kelas data atau konsep yang bertujuan untuk memprediksi atau memperkirakan kelas dari suatu objek dimana kelasnya belum diketahui (Han et al., 2012). Metode klasifikasi terdiri dari *Naive Bayes Classifier*, *Decision Tree*, *K-Nearest Neighbors (KNN)*, *Bayesian Network*, Jaringan Syarat Tiruan (JST), Analisis statistik, Algoritma Genetik, *Rough Sets*, metode berbasis aturan dan SVM.

Berdasarkan cara pelatihan, algoritma klasifikasi dapat dibagi menjadi dua macam yaitu *eager learner* dan *lazy learner*. Pada *eager learner* dilakukan proses pelatihan/pembelajaran pada data latih agar dapat memetakan dengan benar setiap vektor masukan ke label kelas keluarannya sehingga diakhir proses pelatihan, model sudah dapat memetakan data uji dengan benar. Proses prediksi menggunakan model yang tersimpan dan tidak melibatkan data latih sehingga proses prediksi berjalan dengan cepat, tetapi proses pelatihannya memakan waktu lama. Algoritma yang termasuk *eager learner* yaitu Jaringan Syaraf Tiruan, *Decision Tree, Bayesian*, *Support Vector Machine.* Sedangkan *lazy learner* hanya sedikit melakukan pelatihan bahkan tidak. Hal ini menyebabkan proses prediksi menjadi lama karena model harus membaca semua data latih agar dapat memberikan keluaran dengan benar. Kelebihan algoritma ini proses pelatihan berjalan dengan cepat. Algoritma

yang termasuk kategori ini antara lain *K-Nearest Neighbous (KNN), Fuzzy K-Nearest Neighbour*, Regresi Linear, dll (Prasetyo, 2012).

2.2.3 Pengukuran Kinerja Klasifikasi

Proses prediksi yang dilakukan diharapkan mampu melakukan klasifikasi semua data set dengan benar, tetapi tidak dapat dipungkiri jika kinerja suatu sistem tidak 100% benar sehingga perlu dilakukan pengukuran kinerja klasifikasi dengan menggunakan matriks konfusi (*confusion matrix*). Matriks konfusi merupakan tabel pencatat hasil kerja klasifikasi.

Tabel 2.2 Matriks konfusi untuk klasifikasi dua kelas (Prasetyo, 2012)

f_{ij}		Kelas hasil prediksi *(j)*	
		Kelas = 1	Kelas = 0
Kelas asli *(i)*	Kelas = 1	f_{11}	f_{10}
	Kelas = 0	f_{01}	f_{00}

Tabel diatas merupakan contoh matriks konfusi yang melakukan klasifikasi dua kelas yaitu kelas 0 dan 1. Setiap sel f_{ij} menyatakan jumlah data/record kelas *i* yang hasil prediksinya masuk ke kelas *j*. Misal sel f_{11} adalah jumlah data dalam kelas 1 yang secara benar dipetakan ke kelas 1 dan f_{10} adalah data dalam kelas 1 yang dipetakan secara salah ke kelas 0. Jadi berdasarkan tabel diatas jumlah data yang diprediksi secara benar yaitu $f_{11}+f_{00}$

dan data yang diklasifikasi secara salah yaitu $f_{10}+f_{01}$. Kuantitas matriks konfusi dibagi menjadi dua yaitu akurasi dan laju eror atau kesalahan prediksi (Prasetyo, 2012).

Rumus untuk menghitung akurasi :

$$\text{Akurasi}=\frac{\text{Jumlah data diprediksi benar}}{\text{Jumlah prediksi yang dilakukan}}=\frac{f_{11}+f_{00}}{f_{11}+f_{10}+f_{01}+f_{00}} \quad (1)$$

Rumus untuk menghitung laju eror :

$$\text{Laju eror}=\frac{\text{Jumlah data diprediksi salah}}{\text{Jumlah prediksi yang dilakukan}}=\frac{f_{10}+f_{01}}{f_{11}+f_{10}+f_{01}+f_{00}} \quad (2)$$

2.2.4 Bayesian Network (BN)

Bayesian Network atau *Bayesian Belief Network* merupakan suatu metode pemodelan data berbasis probabilitas yang merepresentasikan suatu himpunan variabel dan atribut yang saling berkorespondensi atau berhubungan melalui DAG (*Directed Acyclic Graph*). Bayesian Network memiliki dua tugas yaitu pembelajaran melalui DAG dan struktur dari *bayesian network* berupa jaringan (Friedman et al., 1997).

Bayesian network didasarkan pada Teorema Bayes yaitu *conditional probability* (peluang bersyarat) yang dinotasikan dengan P(A|B) artinya peluang keadaan A jika keadaan B telah terjadi. Berbeda dari *naive bayes* yang mengabaikan hubungan antar atribut atau variabel, pada *bayesian network* antar variabel atau atribut bisa saling dependent atau berhubungan Rumus Teorema Bayes yaitu:

$$P(A|B)=\frac{P(B|A)P(A)}{P(B)} \quad (3)$$

dimana :

P(A|B) = disebut juga *posterior probability*, yaitu peluang A terjadi setelah B terjadi.

P(B∩A) = peluang B dan A terjadi bersamaan

P(B|A) = disebut juga *likelihood*, yaitu peluang B terjadi setelah A terjadi.

P(A) = disebut juga prior, yaitu peluang kejadian A

P(B) = peluang kejadian B

Bayesian network digambarkan seperti graf yang terdiri dari simpul (node) dan busur (*arc*). Node menunjukkan variabel atau atribut beserta nilai probabilitasnya dan busur menunjukkan hubungan antar simpul. Adapun langkah-langkah untuk menerapkan *bayesian network* yaitu (Meigarani et al., n.d.) :

1. Membangun struktur *bayesian network*
2. Menentukan parameter
3. Membuat *Conditional Probability Table* (CPT)
4. Membuat *Joint Probability Distribution* (JPD), untuk menghitung *Joint Probability Distribution* adalah mengalikan nilai *Conditional Probability* dengan *Prior Probability*.
5. Menghitung *Posterior Probabilistik*, didapatkan dari hasil JPD yang telah diperoleh.

6. Inferensi Probabilistik yaitu penelusuran yang dilakukan berdasarkan variabel input yang diberikan pengguna sehingga menghasilkan suatu nilai probabilitas.

2.2.5 K-Nearest Neighbors (K-NN)

Algoritma *K-Nearest Neighbors* merupakan algoritma yang melakukan klasifikasi berdasarkan kedekatan lokasi (jarak) suatu data dengan data yang lain. Nilai K pada K-NN merupakan k-data terdekat dari data uji atau jumlah tetangga terdekat dari data uji (Prasetyo, 2012). Algoritma prediksi K-NN sebagai berikut:

1. Sebuah data uji z= (x',y'), dimana x' adalah atribut data uji sedang y'adalah label kelas data uji yang belum diketahui.
2. Hitung jarak d(x',x) yaitu jarak diantara data uji z ke setiap atribut/vektor data latih, simpan dalam D. Untuk menghitung jarak menggunakan rumus jarak Euclidan, yaitu:

   $$d(x', y') = \sqrt{\sum_{i=1}^{n}(x'_i - y'_i)^2} \quad (4)$$

3. Pilih D_z € D, yaitu K tetangga terdekat dari z
4. $y' = \arg\max \sum(x_i, y_i) \in D_z\, I(v = yi)$

BAB III
METODOLOGI PENELITIAN

Penulis menggunakan beberapa metode penelitian dalam penelitian ini. Adapun metode yang digunakan adalah sebagai berikut:

3.1 Metode Pengumpulan Data

Pengumpulan data dilakukan langsung dilapangan yaitu data alumni mahasiswa UNWIDHA Klaten dengan tahun lulus 2010-2015 terdiri dari beberapa variabel antara lain nilai Indeks Prestasi (IP) dua semester pertama, nilai ujian nasional (UN) yang terdiri dari nilai matematika, bahasa indonesia, bahasa inggris dan ekonomi. Selain itu juga terdapat variabel jurusan sekolah, lulusan sekolah, jalur penerimaan masuk perguruan tinggi dan hasil tes masuk.

3.2 Metode Pengembangan Perangkat Lunak

Pengembangan perangkat lunak dalam penelitian ini dilakukan dengan langkah-langkah sebagai berikut:

1. Analisis, berisi informasi tentang aplikasi atau sistem yang akan dikembangkan.
2. Perancangan/desain sistem, berisi gambaran bentuk sistem yang akan dikembangkan.
3. Pengkodean yaitu proses penulisan *source code* program yang dikembangkan dengan menggunakan bahasa pemrograman.

4. Pengujian perangkat lunak, yaitu proses pengujian terhadap sistem yang dibuat, apakah telah berjalan dengan baik atau belum.

3.3 Bahan Penelitian

Dalam penelitian yang dilakukan menggunakan data mahasiswa yang sudah lulus dengan tahun lulus 2010-2015. Data sampel yang digunakan berasal dari program studi Manajemen yang mewakili jurusan eksak dan program studi Teknik Informatika yang mewakili jurusan non eksak. Data keseluruhan berjumlah 363 *record* yang terdiri dari 194 *record* data dari program studi Teknik Informatika dan 169 record data dari program studi Manajemen.

3.4 Alat Penelitian

3.4.1. Kebutuhan Perangkat Keras

Kebutuhan perangkat keras yang digunakan dalam pembuatan aplikasi ini yaitu sebuah PC *(Personal Computer)* atau laptop dengan spesifikasi:

a. Processor Intel Core i3
b. RAM 1 GB
c. Media penyimpan atau hard disk sebesar 40 GB
d. Perangkat standar input dan output

3.4.2 Kebutuhan Perangkat Lunak

Perangkat lunak yang digunakan dalam penelitian ini antara lain:

a. Microsoft Visual Studio 2010 dengan bahasa pemrograman Visual C#.NET
b. Database Microsoft Access 2010
c. Sistem Operasi Windows XP

3.5 Tahapan Penelitian

Ada beberapa tahapan yang dilakukan dalam penelitian ini antara lain:

1. Mengumpulkan data yaitu data mahasiswa dan data alumni untuk memprediksi masa studi mahasiswa. Data yang diperoleh sebagian dalam bentuk Microsoft Excel dan sebagian lagi dalam bentuk arsip.
2. Melakukan tahapan *preprocessing* data yang terdiri dari:
   a. Pembersihan data yaitu data yang nilai atributnya hilang atau kosong akan dihapus. Proses ini dilakukan diluar sistem, maksudnya sebelum dilakukan proses prediksi oleh sistem sudah dilakukan tahap *preprocessing* yaitu pembersihan data.
   b. Transformasi data, mengubah format jurusan sekolah untuk SMK menjadi IPA atau IPS. Seperti jurusan teknik mesin diubah menjadi IPA, jurusan bisnis & manajemen menjadi IPS. Selain itu mengubah data asal sekolah yaitu SMEA, MAN dan SMU menjadi SMA.
   c. Melakukan pengelompokan nilai atribut atau variabel yang digunakan yang dapat dilihat pada tabel 3.1 dibawah ini.

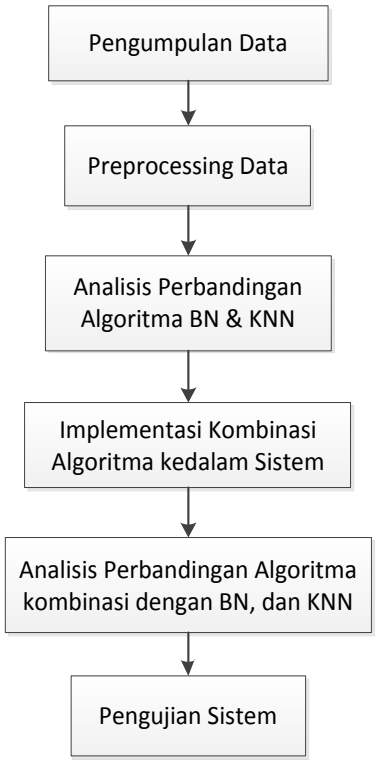
Tabel 3.1 Klasifikasi Variabel Data Set

No	Atribut	Klasifikasi	Kode
1	IP	- < 2 - >= 2 sampai < 2.5 - >= 2.5 sampai < 3 - >= 3 sampai < 3.5 - >= 3.5	-
2.	Jurusan Sekolah	- IPA - IPS - Bahasa	-
3.	Jalur Masuk	- Raport - Test	-
4.	Hasil Tes Masuk	- < 30 - >= 30 sampai < 35 - >= 35 sampai < 40 - >= 40 sampai < 45 - >= 45	-
5.	Nilai UN	- < 5 - >= 5 sampai < 6 - >= 6 sampai < 7 - >= 7 sampai < 8 - >= 8	-
6.	Masa Studi	- < 4 - >= 4 sampai < 4.5 - >= 4.5 sampai < 5 - >= 5 sampai < 5.5 - >= 5.5 sampai < 6 - >= 6	1 2 3 4 5 6

3. Menganalisa perbandingan algoritma *bayesian network* dan *k-nearest neighbors* berdasarkan akurasi dan waktu kecepatan komputasi.
4. Mengkombinasi kedua algoritma diatas dengan mengimplementasikannya kedalam sistem atau aplikasi. Sebelum diimplementasikan kedalam sistem,

terlebih dahulu algoritma kombinasi tersebut diujikan dengan melakukan perhitungan secara manual.

5. Menganalisa perbandingan dari hasil pengujian algoritma kombinasi yang telah dikembangkan dengan *bayesian networks*, dan *k-nearest neighbors*.
6. Melakukan proses pengujian sistem dengan mengukur kinerja klasifikasi pada sistem yaitu berdasarkan tingkat akurasi yaitu seberapa banyak jumlah data yang diprediksi secara benar, dan berdasarkan waktu komputasi yaitu waktu yang dibutuhkan sistem dalam melakukan pengolahan data.



Gambar 3.1 Tahapan Penelitian

3.6 Kendala Penelitian

Dalam melakukan penelitian ini terdapat beberapa kendala yang dihadapi antara lain:

1. Data yang diperoleh selama penelitian sebagian masih berupa dokumen teks jadi belum tersimpan dalam sistem komputer. Setelah semua dokumen terkumpul baru disimpan dalam bentuk excel.
2. Data dengan varibel jalur masuk dan hasil tes masuk tidak semua tersedia karena data tersebut berada dibagian terpisah yaitu bagian penerimaan mahasiswa baru yang belum melakukan pengarsipan data secara komputerisasi, sehingga jumlah data yang digunakan dalam pengolahan data semakin berkurang.

3.7 Algoritma Kombinasi *Bayesian Network* dan *K-Nearest Neighbors*

Pada algoritma kombinasi *Bayesian Network* dan KNN dilakukan dengan mencari nilai posterior probabilistik yang diklasifikasikan pada masing-masing kelas *output*/keluaran P(x|Ci). Kemudian data yang memiliki nilai rata-rata posterior probabilistik > α (dimana α merupakan nilai probabilitas yang diberikan oleh pengguna sistem) akan disimpan untuk digunakan pada proses klasifikasi berikutnya. Jika nilai yang diberikan pengguna tidak ditemukan maka pengguna memasukkan nilai yang lain. Jika data yang disimpan hanya diperoleh satu kelas, maka proses prediksi dengan klasifikasi selesai dan hasil prediksi termasuk kedalam klasifikasi kelas keluaran tersebut. Tetapi jika data yang tersimpan lebih dari 1 kelas dilakukan perhitungan

dengan algoritma KNN yaitu mencari jarak terpendek antara variabel pada data uji dengan data latih.

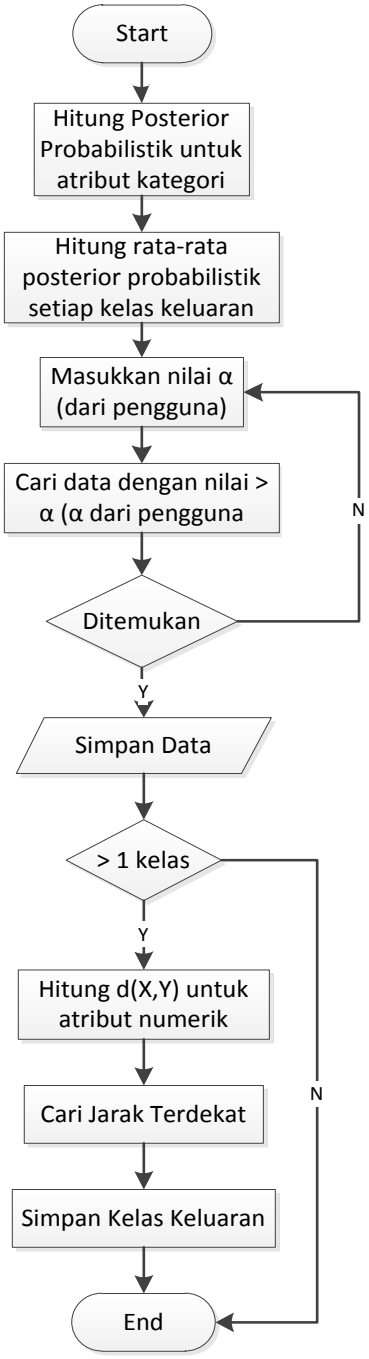
Berikut penjelasan alur algoritma kombinasi *Bayesian Network* dan KNN*:*

1. Langkah pertama adalah menghitung posterior probabilistik untuk masing-masing atribut kategori (lulusan sekolah, jurusan sekolah dan jalur masuk perguruan tinggi). Rumus yang digunakan untuk mencari posterior probabilistik merupakan rumus pada algoritma *bayesian network*.
2. Cari nilai rata-rata posterior probabilistik setiap kelas keluaran. Kemudian simpan data yang mempunyai nilai posterior probabilitas > α (α merupakan nilai masukan dari pengguna yaitu 0-1). Setiap kelas keluaran dicari nilai rata-rata dari hasil pada langkah satu diatas. Nilai α merupakan bilangan desimal antara 0 – 1. Hal dilakukan untuk memfilter data yang akan digunakan pada proses selanjutnya sehingga tidak semua data digunakan. Langkah kedua ini merupakan langkah baru yang dapat dikombinasikan dengan langkah pada algoritma *bayesian network*.
3. Jika data yang memiliki nilai > α tersebut tidak ditemukan pada data pelatihan maka pengguna diminta memasukkan nilai α lain yang lebih kecil dari nilai sebelumnya. Hal ini dapat terjadi jika nilai α yang diberikan lebih besar dari hasil pada langkah kedua sehingga diperlukan nilai α yang lebih kecil lagi. Jika tidak ditemukan juga, langkah ini akan terus dilakukan sampai kondisi pada langkah ini terpenuhi.
4. Jika data yang tersimpan pada langkah kedua hanya mempunyai satu kelas keluaran proses prediksi selesai. Tetapi jika memiliki lebih dari satu kelas,

hitung jarak d(X,Y) antara data uji dengan data latih untuk atribut numerik (IP semester 1, IP semester 2, hasil tes masuk dan nilai UN). Langkah ini merupakan rumus mencari jarak pada algoritma KNN menggunakan rumus jarak *Euclidan* yaitu $d(x', y') = \sqrt{\sum_{i=1}^{n}(x'_i - y'_i)^2}$.

5. Cari jarak terdekat antara data pelatihan dengan data uji yang diperoleh pada langkah keempat. Kelas keluaran yang merupakan jarak terdekat tersebut merupakan hasil prediksi masa studi yang ingin dicari. Langkah ini merupakan langkah terakhir dan proses prediksi selesai.

Untuk *flowchart* dari alur algoritma kombinasi *bayesian network* dan KNN dapat dilihat pada gambar 3.2 berikut ini.



Gambar 3.2 Flowchart Algoritma Kombinasi *Bayesian Network* dan KNN