**BAB 2**

**TINJAUAN PUSTAKA DAN DASAR TEORI**

* 1. **Tinjauan Pustaka**

Rochmana, Lintang Suci (2019) melakukan prediksi penentuan kelulusan mahasiswa tepat waktu dengan mengunakan algoritma *Naive Bayes* dengan atribut jurusan asal sekolah, indeks prestasi semester 1, indeks prestasi kumulatif sampai semester 4, total sks yang ditempuh sampai semester 4, jumlah nilai D (dalam sks) dan jumlah nilai E (dalam sks). Kelas targetnya yaitu tepat waktu dengan 2 klasifikasi ya dan tidak dan diperoleh tingat akurasi sebesar 70%.

Masyithoh (2014) menggunakan *Bayes Network* dalam pengembangan model untuk memprediksi pembelian *smartphone*. Penelitian ini berhasil mengembangkan model yang dapat memprediksi kecenderungan factor-faktor yang mempengaruhi keputusan dan preferensi seorang untuk membeli produk *smartphone* dan dapat mengakomodasi ketidakpastian sehingga dapat menjadi pertimbangan produsen dalam meluncurkan produk *smartphone*. Dalam memodelkan factor-faktor tersebut diadopsi beberapa variable dari teori *technology acceptance model* (TAM), yaitu suatu teori yang cocok untuk mengukur dan menganalisis penerimaan suatu teknologi. Data empiris didapatkan dari 331 responden dengan metode kuesioner. Data diolah menggunakan *path analysis,* hasil penelitian menunjukkan bahwa model prediksi pembelian *smartphone* dibuat dengan pendekatan *Bayes Network* dan mengadopsi teori *technology acceptance model* (TAM) mempunyai akurasi sebesar 98,9%.

Chotimah, Chusnul. (2017) melakukan klasifikasi penyelesaian studi mahasiswa berbasis *Bayes Network* dengan parameter input menggunakan atribut data mahasiswa yang meliputi indeks prestasi kumulatif, jenis kelamin, asal suku, status pernikahan, asal skolah dan jurusan sekolah. Target keluaran system berupa prediksi penyelesaian studi dan diperoleh kesimpulan bahwa dengan sampel data sebanyak 128 menghasilkan nilai akurasi sebesar 71,09%.

Pramitaningrum (2014) melakukan penelitian untuk mengembangkan model prediksi jumlah sampah menggunakan *Bayes Network* yang bertujuan memonitor jumlah sampah di Kota Yogyakarta. Pengembangan model tidak hanya mempertimbangkan data historis tetapi juga mempertimbangkan keterkaitan faktor-faktor lain yang mempengaruhi jumlah sampah seperti jumlah penduduk, pertumbuhan ekonomi dan lainnya. Dalam penelitian ini, dilakukan dengan mengkombinasikan model *holt’s* dan *Bayes Network*. Hasil penelitian model memiliki tingkat akurasi sebesar 67% serta mampu mengikuti pola pergerakan data aktual.  
 Selanjutnya skripsi yang dibuat adalah implementasi metode klasifikasi *Bayesian Network* dalam memprediksi penentuan kelulusan mahasiswa tepat waktu dengan kriteria yang digunakan adalah indeks prestasi semester 1, indeks prestasi kumulatif sampai semester 4, total sks yang ditempuh sampai semester 4, jurusan asal sekolah, jumlah nilai D dan E dalam sks, status kerja, status pernikahan, pendapatan orang tua. Perbedaan dengan penelitian sebelumnya adalah atribut dan metode yang digunakan yaitu menggunakan *Bayeians Network*. Alasan menggunakan metode ini karena data yang akan dilakukan pengujian adalah data kategorikal, data yang bermula dari data numerik diubah ke dalam bentuk data kategorikal. Data ini diubah ke dalam bentuk kategorikal karena dibutuhkan untuk mengetahui penyebab kenapa mahasiswa tidak bisa lulus tepat waktu, misalkan apakah karena nilai ipknya rendah atau total sksnya kurang atau jumlah nilai D-nya banyak atau mungkin nilai E-nya banyak. Alasan pemilihan atribut dengan memprediksi mahasiswa pada semester 4 supaya mahasiswa dapat melakukan perubahan prestasi akademik untuk mencapai target kelulusan yang diinginkan. Dari penelitian yang telah dilakukan oleh peneliti-peneliti sebelumnya terdapat perbedaan yang bisa dilihat pada tabel.

**Table 2.1 Perbandingan Tinjauan Pustaka**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No | Nama Penulis | Objek Penelitian | Metode/ Teknologi | Atribut |
| 1 | Rochmana, Lintang Suci (2019) | Prediksi kelulusan mahasiswa STMIK Akakom Yogyakarta | *naïve bayes* | jurusan asal sekolah, IP semester 1, IPK semester 1-4, total sks yang ditempuh sampai semester 4, jumlah nilai D (dalam sks) dan jumlah nilai E (dalam sks) |
| 2 | Nugroho, M.F. dan Setyoningsih, W. (2017) | Prediksi kelulusan mahasiswa | *Naive Bayes Classifier* | status pekerjaan dan IPK semester 4 | |
| 3 | Chotimah, Chusnul (2017) | Penyelesaian Studi di Universitas Musamus Merauke | *Bayes Network* | indeks prestasi kumulatif, jenis kelamin, asal suku, status pernikahan, asal skolah dan jurusan sekolah |

**2.1 Tabel Lanjutan**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 4 | Nurina Sari, Betha | Akademik siswa pada pelajaran matematika | *Bayesian Networks* danAlgoritma Klasifikasi *Machine Learning* | Kategori dan *numerik kontinyu*. |

* 1. **Program Studi Teknik Informatika**

Dari laman resmi <https://ti.akakom.ac.id/> dikatakan Program Studi Teknik Informatika (PSTI) STMIK Akakom Yogyakarta merupakan salah satu cabang dari pohon ilmu Informatika dan Komputer, yang turut mengembangkan bidang ilmu teknologi informasi. Adapun penekanannya pada aspek penyimpanan, pengolahan, dan penyampaian informasi secara algoritmik dengan bantuan komputer. Tiga (3) paradigma yang digunakan adalah: paradigma teori, paradigma abstraksi, dan paradigma perancangan. Ketiga paradigma ini saling melengkapi dan kait-mengait antara satu dengan yang lain. Secara umum ketiga paradigma tersebut (teori, abstraksi, perancangan) merupakan prinsip dasar dari masing-masing bidang ilmu yang diwakili matematika, ilmu alam, teknik.

Jumlah alumni Program Studi Teknik Informatika  hingga saat ini mencapai sekitar 1000 mahasiswa, yang telah bekerja pada perbankan, perminyakan, pendidikan, kesehatan, dan bidang-bidang lain. Banyak mahasiswa yang telah bekerja sebelum lulus dari pendidikannya. Hal ini membuktikan bahwa ilmu yang diterima mahasiwa bisa digunakan kapan saja. Beberapa lapangan kerja alumni diantaranya Krakatau Steel, BRI, Pertamina, TOTAL, (SGM), Instansi Negeri (Kejaksaaan, Pemda, PTN), dan Perguruan Tinggi Swasta.

* 1. **Penyelesaian Studi**

Penyelesaian studi dapat didefinisikan sebagai indikator ketepatan penyelesaian studi atau ketidaktepatan penyelesaian studi (Ridwan *et al.,* 2013). Untuk mengetahui waktu penyelesaian studi diambil dari tanggal penetapan kelulusan (3,5 tahun – 4 tahun) dianggap lulus tepat waktu dan lebih dari 4 tahun (8 semester) dianggap tidak tepat waktu. Menurut peraturan akademik, penyelesaian studi mahasiswa memiliki dua kategori yaitu lulus tepat waktu atau tidak tepat waktu (sumber : <https://ti.akakom.ac.id/>).

Keputusan Menteri Pendidikan Tinggi dan Penilaian Hasil Belajar Mahasiswa Bab III Pasal 5 Ayat (1) dinyatakan bahwa beban studi program sarjana sekurang-kurangnya 144 (seratus empat puluh empat) sks dan sebanyak-banyaknya 160 (seratus enam puluh) sks yang dijadwalkan untuk 8 semester dan selama-lamanya 14 semester setelah pendidikan menegah. Dari penjelasan di atas dapat disimpulkan bahwa mahasiswa diharapkan mampu menyelesaikan studi dan memperoleh gelar kesarjanaan dengan masa studi lebih cepat yaitu 8 semester atau 4 tahun, tetapi dalam praktiknya mahasiswa tidak selalu dapat menuntaskan studinya selama waktu normal yang telah ditentukan.

Dalam konteks pembelajaran di perguruan tinggi STMIK Akakom (2020) mengatakan bahwa keberhasilkan program pembelajaran yang diperoleh dari kegiatan belajar bersifat kognitif dan biasanya ditentukan melalui pengukuran dan penilaian. Indeks Prestasi (IP) adalah penilaian keberhasilan studi semester mahasiswa yang dilakukan pada tiap semester. Indeks prestasi kumulatif (IPK) merupakan angka yang menunjukkan keberhasilan mahasiswa secara kumulatif mulai dari semester pertama sampai semester paling akhir ditempuh.

Penyelesaian studi atau kelulusan sebagai *output* akhir dari sebuah perguruan tinggi akan diberikan predikat kelulusan. Dasar dari pemberian predikat adalah indeks prestasi kumulatif (IPK). Sebagai contoh di STMIK Akakom Yogyakarta prediksi kelulusan yang ada sebagai berikut : IPK dengan nilai 2,00-2,75 memiliki predikat memuaskan, IPK dengan nilai 2,76-3,50 memiliki predikat sangat memuaskan, dan IPK dengan nilai 3,51-4,00 memiliki predikat terpuji.

* 1. **Weka**

*WEKA* merupakan sebuah perangkat lunak yang menerapkan berbagai algoritma *machine learning* untuk melakukan beberapa proses yang berkaitan dengan sistem temu kembali informasi atau *data mining*. Beberapa fitur unggulan yang dimiliki oleh *WEKA* salah satunya yaitu:

* *Classification*

Di dalam *WEKA* terdapat banyak algoritma yang mendukung untuk proses klasifikasi sebuah objek serta pengguna dimudahkan dalam melakukan implementasi secara langsung. *User* dapat melakukan *load dataset*, melakukan pemilihan algoritma untuk klasifikasi, kemudian diberikan beberapa representasi data yang mewakili hasil akurasi, tingkat kesalahan dari proses klasifikasi.

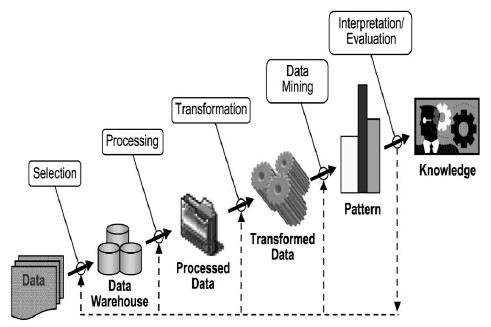
* 1. **Dasar Teori**
     1. ***Data Mining***

*Data mining* adalah kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menemukan keteraturan, pola atau hubungan dalam set data berukuran besar. Keluaran dari data mining ini bisa dipakai untuk memperbaiki pengambilan keputusan di masa depan. (Santosa, 2007).

*Data mining* adalah proses menemukan pengetahuan yang menarik dari jumlah data besar yang disimpan dalam database, data warehouse (gudang data), atau repositori informasi lainnya. (Han & Kamber, 2006).

*Data mining* adalah teknik untuk menemukan dan mendeskripsikan pola-pola yang ada dalam data sebagai sebuah alat untuk membantu menjelaskan data tersebut dan membuat prakiraan dari data itu (Witten & Frank, 2005)

Tahapan yang dilakukan pada proses data mining diawali dari seleksi data dari data sumber ke data target, tahap preprocessing untuk memperbaiki kualitas data, transformasi, data mining serta tahap interpretasi dan evaluasi yang menghasilkan output berupa pengetahuan baru yang diharapkan memberikan kontribusi yang lebih baik. Secara detail dijelaskan sebagai berikut (Fayyad, 1996):



**Gambar 2.1 Tahapan Data Mining**

1. *Data Selection*

*Data Selection* menciptakan himpunan data target atau pemilihan himpunan data, dimana penemuan *(discovery)* akan dilakukan. Pemilihan (seleksi) data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam KDD dimulai.

1. *Pre-processing/ Cleaning*

Sebelum proses *data mining* dapat dilaksanakan, perlu dilakukan proses *cleaning* pada data yang menjadi fokus KDD. Proses *cleaning* antara lain membuang duplikasi data, memeriksa data yang inkonsisten dan memperbaiki kesalahan pada data, seperti kesalahan cetak (*tipografi*).

1. *Transformation*

Proses transformasi adalah proses pencarian fitur-fitur yang berguna untuk mempresentasikan data bergantung kepada *goal* yang ingin dicapai. Proses transformasi dilakukan pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses *data mining*. Proses ini merupakan proses kreatif dan sangat tergantung pada jenis atau pola informasi yang akan dicari dalam basis data.

1. *Data Mining*

*Data mining* dalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik, metode atau algoritma dalam data mining sangat bervariasi.

1. *Interpretation/Evaluation*

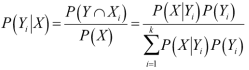
Tahap *interpretation/evaluation* merupakan bagian dari proses KDD yang mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesis yang ada sebelumnya.

* + 1. **Struktur Graf**

Struktur graf bayesian network disebut dengan directed acyclic graph (DAG) yaitu graf berarah yang tidak mengandung siklus. DAG terdiri dari node dan edge. Node merepresentasikan variabel acak dan edge merepresentasikan hubungan kebergantungan langsung (dapat pula diinterpretasikan sebagai pengaruh (sebab–akibat) langsung antara variabel yang dihubungkannya (Ginting, S. L. 2008)

* + 1. **Himpunan Parameter**

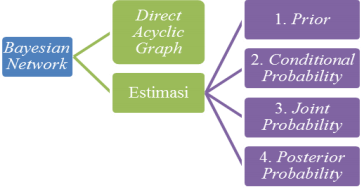
Himpunan parameter mendefinisikan distribusi probabilitas kondisional untuk setiap variabel. Pada bayesian network, node-node berkorespondensi dengan variabel acak. Tiap node diasosiasikan dengan sekumpulan peluang bersyarat, P(Xi|Ai) dimana Xi adalah variabel yang diasosiasikan dengan node dan Ai adalah set dari parent dalam graf. Dalam membangun bayesian network, struktur dibangun dengan pendekatan statistik yang dikenal dengan teorema bayes yaitu conditional probability (peluang bersyarat). Conditional probability yaitu peluang suatu kejadian Y bila diketahui kejadian X telah terjadi, dinotasikan dengan P(Y|X). Rumus teorema bayes :

 (2.1)

*Bayesian network* dapat melakukan pengambilan keputusan (inferensi) probabilistik. Inferensi probabilistik adalah memprediksi nilai variabel yang tidak dapat diketahui secara langsung dengan menggunakan nilai–nilai variabel lain yang telah diketahui. Inferensi probabilistik dapat dilakukan jika terlebih dahulu diperoleh *joint probabillity distribution* (JPD) dari semua variabel yang dimodelkan (Krause, P. J. 1998). JPD adalah probabilitas semua kejadian variabel yang terjadi secara bersamaan. Inferensi probabilistik dapat dilakukan jika *bayesian network* telah dibangun. Dalam kasus ini, hubungan antar variabel dan probabilitas nilai–nilai variabel belum diketahui, oleh karena itu bayesian network dibangun berdasarkan data kejadian, disebut konstruksi bayesian network. Ada dua tahap mengonstruksi *bayesian network*, yaitu : 1. Konstruksi struktur (tahap kualitatif), yaitu mencari keterhubungan antara variabel–variabel yang dimodelkan. 2. Estimasi parameter (tahap kuantitatif), yaitu menghitung nilai–nilai probabilitas.

* + 1. **Membangun Bayesian Network**

Menurut (Ginting, S. L. 2008) membangun bayesian network dibagi dalam dua tahap, lihat pada Gambar 2.3

**

**Gambar 2.3 Bagan Metode *Bayesian Network***

1. Konstruksi Struktur

Struktur *bayesian network* adalah *direct acyclic graph* yang dapat merepresentasikan sebuah pola dari sekumpulan data. Representasinya berbentuk graf yang dilakukan dengan mengindentifikasi konsep– konsep informasi yang relevan terhadap masalah.

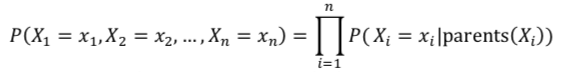
1. Estimasi Parameter

Setelah struktur *bayesian network* terbentuk, parameter (*prior probability*) dan hubungan ketergantungan antara *node* (*conditional probability*) sudah ditentukan berdasar pengetahuan pakar, maka dapat dihitung *joint probability distribution*. Setelah *joint probability distribution* didapat, selanjutnya dihitung *posterior probability*.

* + 1. ***Bayesian Network***

*Bayesian Network* merupakan salah satu *Probabilistic Graphical Model* (PGM) yang sederhana yang dibangun dari teori probabilistik dan teori graf. Teori praobabilistik berhubungan langsung dengan data sedangkan teori graf berhubungan langsung dengan bentuk representasi yang ingin didapatkan. *Bayesian Network* dapat merepresentasikan hubungan sebab akibat diantara variabel-variabel yang terdapat pada struktur *Bayesian Network*. Sebagai contoh, sebuah *Bayesian Network* dapat mewakili hubungan probabilistik antara penyakit dan gejala (Heckerman, 1995). *Bayesian Netwok* terdiri dari dua komponen, grafik *Directed Acyclic Graph* (DAG) dan tabel probabilitas bersyarat (*Conditional Probability Table*, CPT) untuk setiap variabel atribut (Han & Kambe, 2001).

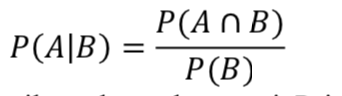
DAG dapat dinotasikan dengan *G = (X,E)* dengan, *X* terdiri dari variabel yang disebut *node* dan *E* adalah pasangan *node* yang dihubungkan dengan garis panah. *Node* pada DAG menggambarkan variabel random, sedangkan garis panah menunjukkan hubungan probabilistik dependensi antar atribut. Garis panah hanya menuju ke satu arah dan tidak berputar kembali menuju *node* asalnya. Jika terdapat garis dari *node* variabel atribut *Xj* menuju *node Xk*, dapat dikatakan bahwa *node Xj* adalah orangtua (*parent*) dari *node Xk* dan *node Xk* adalah anak/turunan (*child*) dari node *Xj*. *Parent* didefinisikan sebagai *node* yang dijadikan syarat dan *child* adalah *node* yang diberikan syarat. Sehingga dapat diperoleh rumus untuk fungsi kepadatan probabilitas (PDF) sebagai berikut:

 (2.2)

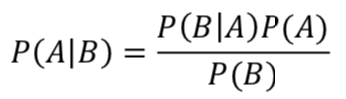
dengan probabilitas *node* turunan hanya tergantung terhadap *node* orangtuanya (Larose, 2006). Dalam suatu *Bayesian Network* dapat dimiliki lebih dari satu *node* keluaran. Selain dapat memperoleh keluaran berupa label kelas, dapat diketahui pula nilai probabilitas untuk setiap kelas.

Jika terdapat suatu parameter/hipotesis pada populasi tersebut, maka parameter tersebut juga mempunyai distribusi yang disebut dengan distribusi posterior.

Distribusi posterior 𝑃(𝐴|𝐵) adalah distribusi bersyarat parameter A diberikan data observasi B dan ditulis secara matematis sebagai berikut :

 (2.3)

Distribusi bersyarat parameter A diberikan data observasi B juga dapat ditentukan dengan menggunakan teorema Bayes berikut :

 (2.4)

dimana,

𝑃(𝐴)  = probabilitas evidence atau bukti

𝑃(𝐵)  = probabilitas prior

𝑃(𝐴|𝐵) = probabilitas bersyarat yang diketahui disebut sebagai likelihood

𝑃(𝐵|𝐴) = probabilitas bersyarat yang akan dicari atau probabilitas posterior