

## BAB II

### TINJAUAN PUSTAKA DAN DASAR TEORI

#### 2.1 Tinjauan Pustaka.

Dalam penelitian ini digunakan beberapa sumber pustaka. Pustaka yang relevan pada penelitian ini di tinjau dari sisi kasus penelitian yang dilakukan adalah mengenai calon mahasiswa baru yang melakukan pendaftaran dan yang melanjutkan ke pendaftaran ulang. Metode yang digunakan dalam penelitian ini adalah *Naïve Bayes*.

Beberapa penelitian yang pernah dilakukan terkait dengan calon mahasiswa baru dengan metode *Naive Bayes* diantaranya adalah yang dilakukan oleh Devi Sugianti (2012), tentang prediksi calon mahasiswa baru di STMIK WIDYA PRATAMA.

Kemudian penelitian oleh Eko Nur Wahyudi (2013). Teknik klasifikasi untuk melihat kecenderungan calon mahasiswa baru dalam memilih jenjang pendidikan program studi perguruan tinggi. Dan penelitian yang dilakukan oleh Tacbir Hendro Pudjiantoro, Faiza Renaldi, dan Age Teogunadi (2011).

Penerapan data mining untuk menganalisa kemungkinan pengunduran diri calon mahasiswa baru. Dengan menggunakan data calon mahasiswa 2007-2009

yang digunakan sebagai data sample untuk pembentukan pohon keputusan, didapatkan tingkat kecocokan hasil prediksi dengan hasil nyata sebesar 61,89 %.

Penelitian ini pernah dilakukan oleh Selvy Meigra, Kusrini dan Emha Taufiq Luthfi (2018) yaitu tentang prediksi heregristrasi calon mahasiswa baru menggunakan algoritma *Naïve Bayes* dengan atribut pendapatan ayah, nilai UN, minat studi, gelombang pendaftaran, jurusan dan status.

Pada penelitian ini, sistem akan melakukan klasifikasi terhadap data calon mahasiswa baru yang didapat dari bagian humas dan marketing STMIK AKAKOM ke dalam dua katagori yaitu mahasiswa yang melakukan pendaftaran ulang dan mahasiswa yang tidak melakukan pendaftaran ulang selama lima tahun ke belakang. Setelah data diklasifikasi kemudian dicari prediksi dari setiap kelas berdasarkan prodi.

**Tabel 2.1 Tabel Perbandingan Penelitian**

Penulis	Objek	Metode	Hasil
Devi Sugianti (2012)	Calon Mahasiswa baru STMIK WIDYA PRATAMA.	Bayesian Classification	Aplikasi ini mampu memprediksi seberapa besar calon mahasiswa baru yang akan registrasi dan heregristrasi.
Eko Nur Wahyudi (2013)	Calon Mahasiswa baru dan Penjurusan PTS lingkungan Kopertiswilayah IV Jawa Tengah	Metode Data Maining, Klasifikasi	Hasil Klasifikasi menunjukkan bahwa bidang ilmu dan program studi tertentu memang didominasi oleh jenis kelamin tertentu seperti laki-laki pada bidang ilmu ekonomi, computer dan hukum, sedangkan perempuan mendominasi pada bidang ilmu kependidikan dan kesehatan.
Tacbir Hendro Pudjiantoro, Faiza Renaldi, dan Age Teogunadi (2011)	Calon Mahasiswa baru di UNJANI	Metode Decision Tree	Dengan mennggunakan data calon mahasiswa 2007-2009 yang digunakan sebagai data sample untuk pembentukan pohon keputusan, didapatkan hasilkecocokan prediksi dengan 600 data calon mahasiswa baru 2010 yang di ujikan terdapat tingkat kecocokan hasil prediksi dengan hasil nyata sebesar 61,89 %.

**Tabel 2.1 Tabel Perbandingan Penelitian (lanjutan)**

Penulis	Objek	Metode	Hasil
Selvy Meigra, Kusrini dan Emha Taufiq Luthfi (2018)	Calon Mahasiswa Baru Universitas AMIKOM	Bayesian Classification	Pada penelitian ini attribute yang digunakan untuk memprediksi herregistrasi calon mahasiswa baru adalah pendapatan ayah, nilai UN, minat studi, jurusan, gelombang pendaftaran.peneleitianini menghasilkan keputusan bahwa metode naïve bayes dapat digunakan untuk melakukan prediksi calon mahasiswa baru.
Presetyo Anugroho, Idris Winarno dan Nur Rosyid M	Email Spam	Naïve Bayes Classifier	Aplikasi ini mampu melakukan klasifikasi email termasuk kedalam spam atau tidak spam dengan data traning 120 data yang telah diklasifikasi meghasilkan tingkat presentasi error 2.5 %.
Yang diajukan	Calon Mahasiswa baru STMIK AKAKOM Yogyakarta	Naïve Bayes	Aplikasi ini mampu memprediksi jumlah mahasiswa baru yang registrasi dan yang tidak melanjutkan registrasi guna membantu bagian humas dan admisi STMIK AKAKOM Yogyakarta dalam memberikan gambaran jumlah mahasiswa baru yang mendaftar.

Perbandingan penelitian yang sudah pernah dilakukan seperti ditunjukkan pada

Tabel 2.1

## **2.2 Dasar Teori**

Penelitian ini menggunakan beberapa dasar teori untuk memudahkan dalam penelitian.

### **2.2.1 Data Mining**

Data mining merupakan proses pengekstraksian informasi dari sekumpulan data yang sangat besar melalui penggunaan algoritma dan teknik penarikan dalam bidang statistik, pembelajaran mesin dan sistem manajemen basis data. Data mining adalah proses menganalisa data dari perspektif yang berbeda dan menyimpulkannya menjadi informasi-informasi penting yang dapat dipakai untuk meningkatkan keuntungan, memperkecil biaya pengeluaran atau bahkan keduanya. Definisi lain mengatakan data mining adalah kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menemukan keteraturan, pola atau hubungan dalam data berukuran besar. Dari beberapa definisi di atas dapat ditarik kesimpulan bahwa data mining merupakan proses ataupun kegiatan untuk mengumpulkan data yang berukuran besar kemudian mengekstraksi data tersebut menjadi informasi – informasi yang nantinya dapat digunakan.

### **2.2.2 Prediksi**

Prediksi memiliki kemiripan dengan estimasi dan klasifikasi. Hanya saja, prediksi hasilnya menunjukkan sesuatu yang belum terjadi (mungkin terjadi di

masa depan).

### **2.2.3 Naive Bayes**

Naive Bayes merupakan sebuah pengklasifikasian probabilistik sederhana yang menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari dataset yang diberikan. Algoritma menggunakan teorema Bayes dan mengasumsikan semua atribut independen atau tidak saling ketergantungan yang diberikan oleh nilai pada variabel kelas. Definisi lain mengatakan Naive Bayes merupakan pengklasifikasian dengan metode probabilitas dan statistik yang dikemukakan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya.

Naive Bayes didasarkan pada asumsi penyederhanaan bahwa nilai atribut secara kondisional saling bebas jika diberikan nilai output. Dengan kata lain, diberikan nilai output, probabilitas mengamati secara bersama adalah produk dari probabilitas individu. Keuntungan penggunaan Naive Bayes adalah bahwa metode ini hanya membutuhkan jumlah data pelatihan (Training Data) yang kecil untuk menentukan estimasi parameter yang diperlukan dalam proses pengklasifikasian. Naive Bayes sering bekerja jauh lebih baik dalam kebanyakan situasi dunia nyata yang kompleks dari pada yang diharapkan.

Persamaan dari teorema Bayes adalah:

$$P(H|X) = \frac{P(X|H) \cdot P(H)}{P(X)}$$

(Persamaan 2.1)

Dimana:

X: Data dengan class yang belum diketahui

H: Hipotesis data merupakan suatu class spesifik

$P(H|X)$ : Probabilitas hipotesis H berdasar kondisi X (posteriori probabilitas)

$P(H)$ : Probabilitas hipotesis H (prior probabilitas)

$P(X|H)$ : Probabilitas X berdasarkan kondisi pada hipotesis H

$P(X)$ : Probabilitas X

Untuk menjelaskan metode Naive Bayes, perlu diketahui bahwa proses klasifikasi memerlukan sejumlah petunjuk untuk menentukan kelas apa yang cocok bagi sampel yang dianalisis tersebut. Karena itu, metode Naive Bayes di atas disesuaikan sebagai berikut:

$$P(C|F_1 \dots F_n) = \frac{P(C)P(F_1 \dots F_n|C)}{P(F_1 \dots F_n)}$$

(Persamaan 2.2)

Dimana variabel C merepresentasikan kelas, sementara variabel  $F_1 \dots F_n$  merepresentasikan karakteristik petunjuk yang dibutuhkan untuk melakukan klasifikasi. Maka rumus tersebut menjelaskan bahwa peluang masuknya sampel karakteristik tertentu dalam kelas C (Posterior) adalah peluang munculnya kelas C (sebelum masuknya sampel tersebut, seringkali disebut prior), dikali dengan peluang kemunculan karakteristik-karakteristik sampel pada kelas C (disebut juga likelihood), dibagi dengan peluang kemunculan karakteristik-karakteristik sampel secara global (disebut juga evidence). Karena itu, rumus di atas dapat pula ditulis secara sederhana sebagai berikut:

$$Posterior = \frac{\text{prior x likelihood}}{\text{evidence}}$$

(Persamaan 2.3)