BAB II

LANDASAN TEORI

2.1 DATA MINING

Data mining merupakan proses semi otomatis yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan *machine learning* untuk mengekstraksi dan mengidentifikasi informasi pengetahuan yang berguna dan bermanfaat yang tersimpan di dalam *database* besar [6].

Data mining adalah serangkaian proses untuk menggali nilai tambah berupa informasi yang selama ini tidak diketahui secara manual dari suatu basis data. Informasi yang dihasilkan diperoleh dengan cara mengekstraksi dan mengenali pola yang penting atau menarik dari data yang terdapat dalam basis data [7].

Data mining adalah kegiatan menemukan pola yang menarik dari data dalam jumlah besar, data dapat disimpan dalam *database*, data *warehouse*, atau penyimpanan informasi lainnya [8].

Jadi dapat disimpulkan bahwa *data mining* merupakan suatu proses untuk mencari informasi baru yang berguna dan bermanfaat dalam suatu data lama yang telah tersusun dan terstruktur yang berskala besar dengan berbagai macam pola dan aturan yang telah ditetapkan.

2.1.1 Tahapan *Data Mining*

Istilah *Knowledge Discovery in Database* (KDD) dan *data mining* sering kali digunakan secara bergantian untuk menjelaskan proses penggalian informasi

tersembunyi dalam suatu basis data yang besar. Sebenarnya kedua istilah tersebut memiliki konsep yang berbeda, tetapi berkaitan satu sama lain, dan salah satu tahapan dalam keseluruhan proses KDD adalah *data mining*. Proses KDD secara garis besar dapat dijelaskan sebagai berikut [9] :

a. *Data Selection*

Pemilihan data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam *Knowledge Discovery in Database* (KDD) dimulai. Data hasil seleksi yang akan digunakan untuk proses *data mining*, disimpan dalam suatu berkas terpisah dari basis data operasional.

b. *Cleaning*

Sebelum proses *data mining* dapat dilaksanakan, perlu dilakukan proses *cleaning* pada data yang menjadi fokus *Knowledge Discovery in Database* (KDD). Proses *cleaning* mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data, seperti kesalahan cetak. Selain itu juga dilakukan proses *enrichment*, yaitu proses "memperkaya" data yang sudah ada dengan data atau informasi lain yang relevan dan diperlukan untuk *Knowledge Discovery in Database* (KDD), seperti data atau informasi eksternal lainnya yang diperlukan.

c. Transformasi

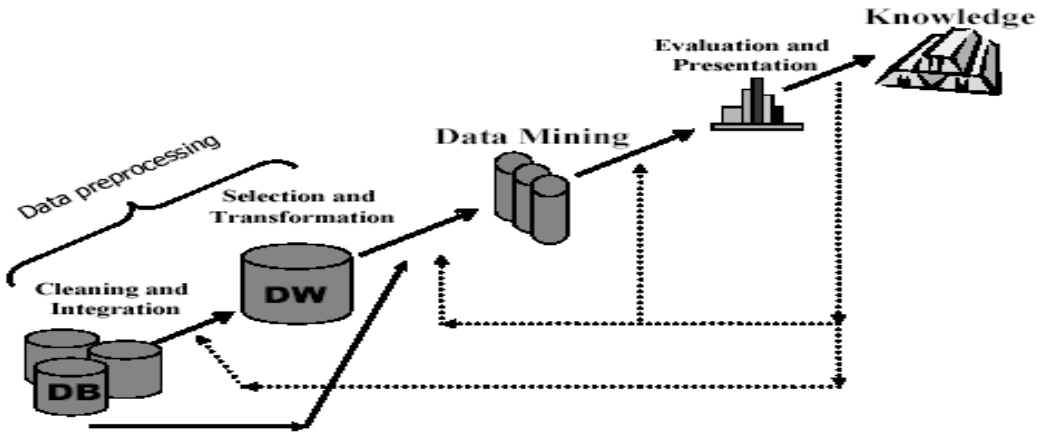
Coding adalah proses transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses *data mining*. Proses *coding* dalam *Knowledge Discovery in Database* (KDD) merupakan proses kreatif dan sangat tergantung pada jenis atau pola informasi yang akan dicari dalam basis data.

d. *Data Mining*

Data Mining adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik-teknik, metode-metode, atau algoritma dalam *data mining* sangat bervariasi. Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses *Knowledge Discovery in Database* (KDD) secara keseluruhan.

e. Interpretasi / Evaluasi

Pola informasi yang dihasilkan dari proses *Data Mining* perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan. Tahap ini merupakan bagian dari proses *Knowledge Diccovery in Database* (KDD) yang disebut *interpretation*. Tahap ini mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesis yang ada sebelumnya.



Gambar 2. 1 Proses *Knowledge Discovery in Database* (KDD) [9]

2.1.2 Pengelompokan *Data Mining*

Data Mining dibagi menjadi beberapa kelompok berdasarkan tugas yang dapat dilakukan, yaitu [9]:

a. Deskripsi

Terkadang peneliti dan analis secara sederhana ingin mencoba mencari cara untuk menggambarkan pola dan kecenderungan yang terdapat dalam data. Sebagai contoh, petugas pengumpulan suara mungkin tidak dapat menemukan keterangan atau fakta bahwa siapa yang tidak cukup profesional akan sedikit didukung dalam pemilihan presiden. Deskripsi dari pola dan kecenderungan sering memberikan kemungkinan penjelasan untuk suatu pola atau kecenderungan.

b. Estimasi

Estimasi hampir sama dengan klasifikasi, kecuali variabel target estimasi lebih ke arah numerik daripada ke arah kategori. Model 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.

c. Prediksi

Prediksi hampir sama dengan klasifikasi dan estimasi, kecuali bahwa dalam prediksi nilai dari hasil akan ada dimasa mendatang. Beberapa metode dan teknik yang digunakan dalam klasifikasi dan estimasi dapat pula digunakan (untuk keadaan yang tepat) untuk prediksi.

d. Klasifikasi

Suatu pengelompokan data di mana data yang digunakan tersebut mempunyai kelas label atau target. Sebagai contoh, penggolongan pendapatan dapat dipisahkan dalam tiga kategori, yaitu pendapatan tinggi, pendapatan sedang, dan pendapatan rendah.

e. *Clustering*

Clustering merupakan pengelompokan *record*, pengamatan, atau memperhatikan dan membentuk kelas objek-objek yang memiliki kemiripan. Kluster adalah kumpulan *record* yang memiliki kemiripan satu dengan yang lainnya dan tidak sama dengan *record* - *record* dalam kluster lain.

f. Asosiasi

Tugas asosiasi dalam *data mining* adalah menemukan atribut yang muncul dalam satu waktu. Dalam dunia bisnis lebih umum disebut analisis keranjang belanja

2.2 KLASIFIKASI

Klasifikasi adalah proses penemuan model atau fungsi yang menggambarkan dan membedakan kelas data atau konsep yang bertujuan agar bisa digunakan untuk memprediksi kelas dari objek yang label kelasnya tidak diketahui [10].

Klasifikasi adalah teknik yang dilakukan untuk memprediksi *class* atau properti dari setiap *instance* data. Model prediksi memungkinkan untuk memprediksi nilai-nilai variabel yang tidak diketahui berdasarkan nilai variabel lainnya. Klasifikasi memetakan data ke dalam kelompok-kelompok kelas yang telah ditetapkan sebelumnya [11].

Klasifikasi adalah proses menemukan model dari *training set* yang membedakan atribut ke dalam kategori atau kelas yang sesuai, model tersebut kemudian digunakan untuk mengklasifikasikan atribut yang kelasnya belum diketahui sebelumnya [12].

Jadi dapat disimpulkan bahwa klasifikasi merupakan suatu cara atau usaha mencari suatu nilai yang belum diketahui berdasarkan atribut ataupun variabel yang telah ditentukan.

Terdapat beberapa algoritma yang digunakan dalam klasifikasi, di antaranya yaitu :

1. *Naïve Bayes.*

   Naive Bayes merupakan sebuah pengklasifikasian probabilistik sederhana yang menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari *dataset* yang diberikan [13].

2. *Decision Tree*

   Decision Tree adalah sebuah struktur yang dapat digunakan untuk mengubah data menjadi pohon keputusan yang akan menghasilkan aturan-aturan keputusan besar menjadi himpunan-himpunan *record* yang lebih kecil dengan menerapkan serangkaian aturan keputusan [14].

3. *Support Vector Machine* (SVM)

   Support Vector Machine (SVM) yaitu sistem pembelajaran yang menggunakan ruang hipotesis berupa fungsi – fungsi linier dalam sebuah fitur yang berdimensi tinggi dan dilatih dengan menggunakan algoritma pembelajaran yang didasarkan pada teori optimasi [15].

4. *Neural Network*

   Jaringan syaraf tiruan merupakan salah satu representasi buatan dari otak manusia yang menyimulasikan cara kerja pada sistem syaraf manusia dalam menjalankan tugas tertentu. Dasar dari pemodelan ini yaitu, kemampuan otak

manusia dalam mengatur sel-sel penyusunnya yang disebut neuron sehingga dapat menjalankan tugas dengan baik. Layaknya sistem syaraf manusia, terdapat banyak neuron yang dimiliki oleh jaringan syaraf tiruan. Neuron-neuron ini tersebar pada beberapa lapisan yaitu, lapisan masukan, lapisan tersembunyi, dan lapisan keluaran [16].

5. *Fuzzy*

   Logika *Fuzzy* merupakan logika bernilai banyak/*multivalued logic* yang mampu mendefinisikan nilai di antara keadaan yang konvensional seperti benar atau salah, ya atau tidak, putih atau hitam dan lain-lain [17].

2.3 BEASISWA

Beasiswa dapat dikatakan sebagai pembiayaan yang tidak bersumber dari pendanaan sendiri atau orang tua, akan tetapi diberikan oleh pemerintah, perusahaan swasta, kedutaan, universitas, serta lembaga pendidik atau peneliti, atau juga kantor tempat bekerja yang karena prestasi seorang karyawan dapat diberikan kesempatan untuk meningkatkan kapasitas sumber daya manusianya melalui pendidikan [18].

Beasiswa adalah program yang bertujuan untuk meningkatkan akses dan pemerataan kesempatan belajar bagi seluruh rakyat Indonesia, mengurangi jumlah pelajar yang putus sekolah atau kuliah karena tidak mampu membayar biaya pendidikan studi, serta meningkatkan prestasi dan motivasi pelajar dalam menempuh pendidikannya[1].

Jadi dapat disimpulkan bahwa beasiswa merupakan suatu program ataupun bantuan dari pemerintah, yayasan, ataupun perusahaan kepada seseorang yang

membutuhkan bantuan dana supaya bisa tetap melanjutkan dan meningkatkan pendidikannya.

2.4 NAÏVE BAYES

Teorema Bayes merupakan teknik prediksi berdasarkan kemungkinan sederhana pada penerapan aturan *Bayes* dengan ketidaktergantungan yang kuat. Naïve Bayes banyak digunakan untuk proses klasifikasi karena *Naïve Bayes* lebih disukai disebabkan kecepatan dan kesederhanaannya [19].

Metode *Naive Bayes Classifier* merupakan sebuah pengklasifikasian probabilistik sederhana yang menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari *dataset* yang diberikan[20].

Naïve Bayes Classifier merupakan sebuah teknik untuk klasifikasi menggunakan 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 [21]. Keuntungan penggunaan *Naïve bayes* adalah bahwa metode ini hanya membutuhkan jumlah data pelatihan (*Training Data*) yang kecil untuk menentukan estimasi paremeter yang diperlukan dalam proses pengklasifikasian. *Naïve bayes* sering bekerja jauh lebih baik dalam kebanyakan situasi dunia nyata yang kompleks daripada yang diharapkan [22]. Selain itu, kelebihan *Naïve bayes* juga dapat menangani data kuantitatif dan data diskrit, kokoh terhadap *noise*, dan dapat menangani nilai yang hilang dengan mengabaikan instansiasi selama perhitungan

estimasi peluang. Persamaan dari teorema *Bayes* [23] dapat dilihat pada Persamaan 2.1 :

$$P(C_i|X) = \frac{P(X|C_i))}{P(X)} \quad \textbf{(2.1)}$$

Dimana :

X : Kriteria suatu kasus berdasarkan masukan

C_i : Kelas solusi pola ke-i, dimana i adalah jumlah label kelas.

$P(C_i\backslash X)$: Probabilitas kemunculan label kelas C_i dengan kriteria masukkan X

$P(X|C_i)$: Probabilitas kriteria masukkan X dengan label kelas C_i.

$P(C_i)$: Probabilitas label kelas C_i

Kinerja sistem klasifikasi menggambarkan seberapa baik sistem dalam mengklasifikasikan data. *Confusion matrix* merupakan salah satu metode yang dapat digunakan untuk mengatur kinerja suatu metode pada klasifikasi. Pada dasarnya *Confusion matrix* mengandung informasi yang membandingkan hasil klasifikasi yang dilakukan oleh sistem dengan hasil klasifikasi yang seharusnya. Pada pengukuran kinerja menggunakan *Confusion matrix*, terdapat 4 (empat) istilah sebagai representasi hasil proses klasifikasi, keempat istilah tersebut adalah *True Positive* (TP), *True Negative* (TN), *False Positive* (FP) dan *False Negative* (FN) [24]. Nilai *True Negative* (TN) merupakan jumlah data negatif yang terdeteksi dengan benar, sedangkan *False Positive* (FP) merupakan data negatif namun terdeteksi sebagai data positif. Sementara itu, *True Positive* (TP) merupakan data positif yang terdeteksi benar. *False* Negative (FN) merupakan kebalikan dari *True Positive*, sehingga data positif, namun terdeteksi sebagai data negatif.

Pada jenis klasifikasi binary yang hanya memiliki 2 keluaran kelas, *Confusion matrix* dapat dilihat pada Tabel 2.1 :

Tabel 2. 1 Tabel *Confusion Matrix*

Kelas	Terklasifikasi Positif	Terklasifikasi Negatif
Positif	TP (True Positif)	FN (False Negative)
Negative	FP (False Positive)	TN (True Negative)

2.5 SELEKSI FITUR *INFORMATION GAIN*

Information Gain (IG) merupakan suatu pengukuran yang dilakukan untuk melakukan seleksi terhadap atribut-atribut sehingga dapat disimpulkan atribut apa saja yang akan digunakan. *Information Gain* menggunakan entropy untuk menentukan atribut terbaik. Entropy merupakan ukuran ketidakpastian dimana semakin tinggi entropy, maka semakin tinggi ketidakpastian [25].

$$Entropy(S)=\sum -p_i \times_{ni=1} log_2(p_i) \quad \text{..... (2.2)}$$

S himpunan kasus, n adalah jumlah dari S dan $p\ i$ jumlah sampel untuk kelas i. Setelah nilai entropy dihitung, selanjutnya hitung nilai *Information Gain*, untuk mengetahui nilai gain dari setiap atribut

$$Gain(S,A)=Entropy(S)-\sum |S\ i\ ||S\ |n\ i=1\times Entropy(S\ i)$$

)**.....(2.3)**

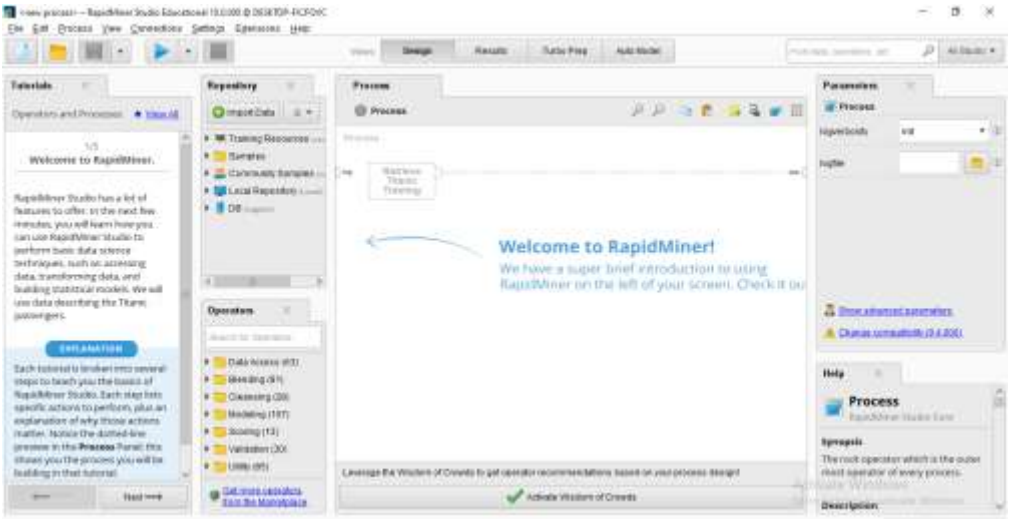
S himpunan kasus, A merupakan atribut, n jumlah atribut dari A , $|S\ i\ |$jumlah kasus untuk partisi ke-i dan $|S\ |$jumlah kasus dari S

2.6 RAPIDMINER

RapidMiner adalah koleksi dari algoritma *learning machine* yang digunakan untuk tugas-tugas *data mining. RapidMiner* berisi *tool* untuk data *pre-processing*, klasifikasi, regresi, *clustering, rule association*, dan memvisualisasikan data tersebut menjadi mudah untuk dapat dipahami [26].

RapidMiner merupakan *software*/perangkat lunak untuk pengolahan data. Dengan menggunakan prinsip dan algoritma *data mining, RapidMiner* mengekstrak pola-pola dari *dataset* yang besar dengan mengkombinasikan metode statistika, kecerdasan buatan dan *database* [27].

RapidMiner memudahkan penggunanya dalam melakukan perhitungan data yang sangat banyak dengan menggunakan operator-operator. Operator ini berfungsi untuk memodifikasi data. Data dihubungkan dengan *node-node* pada operator kemudian kita hanya tinggal menghubungkannya ke *node* hasil untuk melihat hasilnya. Hasil yang diperlihatkan *RapidMiner* pun dapat ditampilkan secara visual dengan grafik. Menjadikan *RapidMiner* adalah salah satu *software* pilihan untuk melakukan ekstraksi data dengan metode-metode *data mining* [28]. Tampilan halaman utama aplikasi *RapidMiner* dapat dilihat pada gambar 2.2 :



Gambar 2. 2 Tampilan Halaman Awal *RapidMiner*

2.7 PENELITIAN SEJENIS

Penulis memasukkan beberapa kajian dari penulis-penulis yang terlebih dahulu telah melakukan dan menyelesaikan penelitian untuk menjadi perbandingan serta acuan bagi penulisan karya ilmiah berjudul "Penerapan *Data Mining* Untuk Klasifikasi Kelayakan Beasiswa Menggunakan Metode *Naïve Bayes* Dengan Seleksi Fitur *Information Gain* (Studi Kasus : SDN 47/IV Kota Jambi)". Berikut adalah kajian penelitian sejenis yang dapat dilihat pada tabel 2.2:

Tabel 2. 2 Penelitian Sejenis

No	Peneliti	Judul	Masalah	Metode	Hasil Penelitian
1.	Ar Razi[29].	Klasifikasi Penerima Beasiswa Aceh Carong (Aceh Pintar) Di Universitas Malikussaleh Menggunakan Algoritma KNN (K-Nearest Neighbors) (2022).	Beasiswa Aceh carong merupakan program dari Pemerintah Aceh untuk membantu putra putri Aceh yang ingin melanjutkan pendidikan dengan memperoleh beasiswa, tetapi pemerintah masih menggunakan cara manual mengakibatkan kurang tepatnya penerima beasiswa tersebut.	*KNN (K-Nearest Neighbors)*	Pada Penelitian ini menggunakan metode KNN didapatkan hasil akurasi sebesar 82% dengan 4 atribut berupa nilai IPK, Semester berjalan, Penghasilan Orang Tua, dan Jumlah Tanggungan Orang tua.
2.	Gagan Suganda, Marsani Asfi, Ridho Taufiq Subagio, Ricky Perdana Kusuma[30]	Penentuan Penerima Bantuan Beasiswa Kartu Indonesia Pintar (KIP) Kuliah Menggunakan Naive Bayes Classifier (2022).	Proses seleksi calon mahasiswa penerima bantuan KIP kuliah yang masih dilakukan secara manual yaitu dengan pengumpulan berkas secara langsung sebagai persyaratan yang telah ditentukan. Maka dari itu dibutuhkan pengambilan keputusan dikarenakan jumlah yang mendaftar	*Naive Bayes*	Jumlah data yang digunakan pada penelitian ini berjumlah 100 data yang dibagi menjadi 90 data *training* dan 10 data *testing*. Atribut yang digunakan yaitu penghasilan ayah dan ibu, status ayah dan ibu, jumlah tanggungan, kepemilikan rumah dan MCK, sumber air dan listrik, luas tanah dan bangunan, prestasi, serta hasil tes ujian yang telah diikuti. Hasil akurasi yang didapatkan pada penelitian ini adalah 88,21%.

			selalu melebihi kouta yang telah ditetapkan dan juga proses pengolahan data membutuhkan waktu yang lama.		
3.	Lutfi Abdullah, Rosmawati Tamin, A. Akhmad Qashlim[31].	Klasifikasi Penerimaan Beasiswa Menggunakan Algoritma Naive Bayes Di Universitas Al Asyariah Mandar Kabupaten Polewali Mandar (2021).	Setiap institusi pendidikan memilik beberapa program kerja yang dapat membantu mahasiswa, maka dibutuhkan penelitian ini untuk membantu pihak kampus dalam menyeleksi mahasiswa yang layak mendapatkan beasiswa agar tepat sasaran.	*Naive Bayes*	Penelitian ini menggunakan metode *naive bayes* yang terdiri dari 13 data. Atribut yang dipakai ada 6 yaitu Nilai IPK, Jarak rumah dari kampus, kendaraan, tempat tinggal, jumlah tanggungan, dan pendapatan orang tua. Hasil yang didapatkan pada penelitian ini yaitu mempunyai akurasi sebesar 88% dari keseluruhan percobaan.
4.	Bayu Maeky Nugroho, Triawan Adi Cahyanto M.Kom[32].	Klasifikasi Penerimaan Beasiswa Menggunakan Algoritma C.45 (Studi Kasus SMA PGRI Cluring Kabupaten Banyuwangi) (2019).	Pihak sekolah membutuhkan sistem agar pemberian program beasiswa PIP dapat tepat sasaran dan diterima oleh siswa yang memang berhak menerimanya.	C4.5	Pada penelitian ini menggunakan data sebanyak 163 data dengan 5 atribut berupa Penghasilan Orang Tua, Penerima Kartu Penjamin Sosial (KPS), Layak Program Indonesia Pintar (PIP), Jumlah Saudara Kandung dan Penerima Kartu Indonesia Pintar (KIP). Hasil dari penelitian ini didapatkan akurasi sebesar 76,74% dan presisi 78,94% dengan menggunakan 163 data.

5.	Fadila Nur Indrasari[33].	Klasifikasi Penerimaan Beasiswa Menggunakan Algoritma Naive Bayes Classifier (2018).	Kurang tepatnya sasaran penerima beasiswa di akibatkan kesulitan dalam menyeleksi mahasiswa dan kriteria yang banyak. Akibatnya mahasiswa yang seharusnya layak mendapatkan beasiswa malah tidak dapat menerimanya.	*Naive Bayes*	Penelitian menggunakan algoritma *Naive Bayes Classifier* untuk mengatasi permasalahan tersebut. Dengan algoritma *Naive Bayes* ini dapat dihitung nilai probabilitas yang digunakan untuk menentukan apakah mahasiswa tersebut berhak menerima beasiswa ataupun tidak.. Atribut yang digunakan ada 7 yaitu Nama, NPM, Prodi, Semester, IPK, Alamat, Jumlah Saudara, Pekerjaan Orang Tua . Data yang dipakai merupakan data kuisioner tahun 2016-2017 dengan jumlah data sebanyak 100 data. Hasil akurasi yang didapatkan pada penelitian ini yaitu sebesar 87,50%.

Berdasarkan penelitian sejenis di atas, dapat dijelaskan bahwa metode *Naive Bayes* memiliki kinerja yang lebih baik dibandingkan metode lain dengan akurasi di atas 87%. Penelitian ini memiliki perbedaan dengan penelitian lainnya yaitu terletak pada jumlah data, atribut dan metode seleksi fitur. Jumlah data siswa pada penelitian ini ada 969 data. Atribut yang dipakai pada penelitian ini ada 15 yaitu Nama, Jenis Kelamin, Kategori Usia Ayah, Jenjang Pendidikan Ayah, Pekerjaan Ayah, Penghasilan Ayah, Jenjang Pendidikan Ibu, Pekerjaan Ibu, Penghasilan Ibu, Jumlah Tanggungan, Jarak Rumah ke Sekolah, Alat Transportasi, Penerima KPS, Penerima KIP dan Layak PIP. Penelitian ini menggunakan seleksi atribut *Information Gain* untuk menyeleksi atribut yang akan

diterapkan dalam perhitungan supaya hasil klasifikasi yang didapatkan lebih akurat. Penulis juga akan menggunakan alat bantu berupa *RapidMiner* untuk membantu proses klasifikasi beasiswa pada SDN 47/IV Kota Jambi. Hasil yang diperoleh dapat dijadikan acuan oleh pihak sekolah dalam menentukan kelayakan penerima beasiswa.