

METODE NAIVE BAYES CLASSIFER DALAM PENENTUAN PENERIMA BEASISWA BIDIKMISI DI UNIVERSITAS NEGERI MEDAN

Said Iskandar¹, Nice R Refisis², Billroy A Ginting³

¹saidiskandar@unimed.ac.id

^{1,2,3} Universitas Negeri Medan

Jl. Williem Iskandar, Psr V, Medan, Sumatera Utara

Abstrak—Beasiswa diberikan kepada mahasiswa dengan tujuan mengurangi jumlah mahasiswa yang putus kuliah, karena tidak mampu membiayai pendidikan. Namun untuk mendapatkan beasiswa mahasiswa harus memenuhi syarat yang telah ditetapkan. Oleh karena jumlah mahasiswa yang mengajukan permohonan harus memiliki kriteria penilaian yang ditentukan oleh universitas. Penerapan Metode Algoritma Naive Bayes Classifier pada evaluasi kinerja akademik mahasiswa dapat membantu memberikan rekomendasi penerima beasiswa. Naive Bayes adalah suatu metode klasifikasi dalam data mining dengan menggunakan metode probabilitas dan statistik. Data mining adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan machine learning untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai database besar. Penelitian ini adalah Studi Kasus yang dilakukan di Jurusan Matematika Angkatan 2019 Universitas Negeri Medan. Data yang diambil adalah pekerjaan orang tua, penghasilan orang tua, jumlah tanggungan, daya listrik (watt), dan nilai ujian nasional. Adapun beberapa tahapan dari proses prosedur dari Klasifikasi Naive Bayes Classifier (NBC) yaitu : Pengumpulan Data, Data Cleaning, Data Transformation, dan Proses Perhitungan Naive Bayes Classifier. Dari hasil penelitian didapat Pengujian pada perbandingan data training dan data testing sebesar 80:20 menghasilkan akurasi tertinggi dengan 79% dan dilihat Metode Naive Bayes Classification digunakan untuk mengklasifikasikan Penerima beasiswa menghasilkan akurasi yang baik.

Keywords— Beasiswa, Naive Bayes Classifier, Data Mining.

Abstract—Scholarships are given to students with the aim of reducing the number of students who drop out of college, because they cannot afford to pay for their education. However, to get a scholarship, students must meet certain conditions. Therefore, the number of students who apply must have the assessment criteria determined by the university. The application of the Naive Bayes Classifier Algorithm in the academic evaluation of students can help provide recommendations for scholarship recipients. Naive Bayes is a classification method in data mining using probability and statistical methods. Data mining is one that uses statistical, mathematical, artificial intelligence, and machine learning techniques to extract and identify useful information and related knowledge from large databases. This research is a case study conducted at the Department of Mathematics Class of 2019 Medan State University. The data taken are parents' occupations, parents' income, number of dependents, electric power (wattage), and national exam scores. There are several stages of the procedure process of the Naive Bayes Classifier (NBC), namely: Data Collection, Data Cleaning, Data Transformation,

and the Naive Bayes Classifier Calculation Process. From the results of the research obtained, the test on the comparison of training data and testing data is 80:20 with the highest accuracy quality with 79% and seen from the Naive Bayes Classification method used to classify good quality scholarship recipients.

Keywords— Scholarship, Naive Bayes Classifier, Data Mining.

1. PENDAHULUAN

Untuk mendapatkan beasiswa mahasiswa harus memenuhi syarat yang telah ditetapkan. Ada beberapa hal kriteria yang ditetapkan untuk mendapatkan beasiswa seperti Pekerjaan orangtua, Penghasilan orangtua, Jumlah tanggungan, jumlah daya listrik(Watt) dan Nilai Ujian Nasional. Oleh sebab itu, tidak semua mahasiswa yang mengajukan permohonan untuk menerima beasiswa dapat dikabulkan. Oleh karena jumlah mahasiswa yang mengajukan permohonan yang banyak dan kriteria penilaian yang banyak pula, maka perlu suatu formula yang dapat membantu memberikan rekomendasi penerima beasiswa.

Penelitian sebelumnya telah dilakukan tentang untuk evaluasi kinerja akademik mahasiswa diantaranya adalah Penerapan Metode Algoritma *Naive Bayes Classifier* pada evaluasi kinerja akademik mahasiswa(Hadi Suyono 2013).

Algoritma *Naive Bayes Classifier NBC* merupakan salah satu algoritma dalam teknik data mining yang menerapkan teori Bayes dalam klasifikasi . Teorema keputusan Bayes adalah pendekatan statistik yang fundamental dalam pengenalan pola (*pattern recognition*). 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. Dan Pengujian pada data mahasiswa angkatan 2005-2009 mining NBC menghasilkan nilai precision, recall, dan accuracy masing-masing 83%, 50%, dan 70%.

Penelitian lainnya telah dilakukan tentang penerimaan beasiswa mahasiswa

STMIK Sinar Nusantara Surakarta diantaranya adalah Penerapan Metode *K-Nearest Neighbor* pada aplikasi penentu penerima beasiswa mahasiswa di STMIK Sinar Nusantara Surakarta (Hendri 2014). Metode *K-Nearest Neighbor* (KNN) merupakan metode klasifikasi yang dapat digunakan untuk memberikan keputusan alternatif dalam menentukan calon penerima beasiswa berdasarkan data-data yang telah diperoleh (data sampel). Dari pengujian yang dilakukan terhadap 22 data sampel yang dijadikan acuan dalam perhitungan *K-Nearest Neighbor* dalam menghasilkan keputusan diperoleh nilai keakuratan sebesar 90, 90% yang mana dalam algoritma K-Nearest Neighbor nilai tersebut termasuk besar karena algoritma K-Nearest Neighbor tidak menggunakan parameter untuk dijadikan acuan melainkan data sampel yang nilainya bervariasi.

Sedangkan *Bayesian Classification* adalah pengklasifikasian statistik yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu class. *Bayesian Classification* didasarkan pada teorema Bayes yang memiliki kemampuan klasifikasi serupa dengan *Decesion Tree* dan *Neural Network*. *Bayesian Classification* terbukti memiliki akurasi dan kecepatan yang tinggi saat diaplikasikan ke dalam database dengan data yang besar . Metode Bayes merupakan pendekatan statistic untuk melakukan inferensi induksi pada persoalan klasifikasi. Pertama kali dibahas terlebih dahulu tentang konsep dasar dan definisi pada Teorema Bayes, kemudian menggunakan teorema ini untuk melakukan klasifikasi dalam Data Mining.

Berdasarkan permasalahan yang sudah dijelaskan, maka penulis tertarik untuk mengadakan penelitian dengan memilih

judul "Penentuan Penerima Beasiswa Bidikmisi dengan Metode Naive Bayes Classifier di Universitas Negeri Medan".

2. BEASISWA BIDIKMISI

Beasiswa bidik misi adalah program bantuan biaya pendidikan yang diberikan Pemerintah melalui Direktorat Jenderal Pendidikan Tinggi (Ditjen Dikti) Kementerian Pendidikan dan Kebudayaan mulai tahun 2010 kepada mahasiswa yang memiliki potensi akademik memadai dan kurang mampu secara ekonomi. Universitas Tanjungpura sebagai salah satu perguruan tinggi negeri yang berada dibawah Kementerian Pendidikan dan Kebudayaan juga mendapat bantuan bidik misi, yang lebih dikenal dengan beasiswa penuh outreaching.

Adapun tujuan beasiswa Bidikmisi secara umum yaitu:

1. Meningkatkan akses dan kesempatan belajar di perguruan tinggi bagi peserta didik yang tidak mampu secara ekonomi dan berpotensi akademik baik.
2. Memberi bantuan biaya pendidikan kepada calon/mahasiswa yang memenuhi kriteria untuk menempuh pendidikan program diploma/sarjana sampai selesai dengan tepat waktu.
3. Meningkatkan prestasi mahasiswa baik pada bidang kurikuler, ko kurikuler maupun ekstra kurikuler.
4. Menjamin keberlangsungan studi mahasiswa sampai selesai dan tepat waktu. (Kemenristekdikti 2015)

Program Bantuan Biaya Pendidikan Bidikmisi yang dimulai peluncuran programnya pada tahun 2010 ini adalah bantuan biaya pendidikan bagi calon mahasiswa tidak mampu secara ekonomi dan memiliki potensi akademik baik untuk menempuh pendidikan di perguruan tinggi pada program studi unggulan sampai lulus tepat waktu. Adapun persyaratan bagi para calon pendaftar beasiswa bidikmisi :

1. Siswa SMA/SMK/MA/MAK atau bentuk lain yang sederajat yang akan lulus pada tahun 2019;

2. Lulusan tahun 2018 yang bukan penerima Bidikmisi dan tidak bertentangan dengan ketentuan penerimaan mahasiswa baru di masing-masing perguruan tinggi;
3. Usia paling tinggi pada saat mendaftar adalah 21 tahun;
4. Tidak mampu secara ekonomi dengan kriteria:
 - a) Siswa penerima Beasiswa Siswa Miskin (BSM);
 - b) Pemegang Kartu Pengaman Sosial (KPS) atau sejenisnya;
 - c) Pendapatan kotor gabungan orangtua/wali (suami istri) sebesar-besarnya Rp3.000.000,00 perbulan. Untuk pekerjaan nonformal/informal pendapatan yang dimaksud adalah rata-rata penghasilan per bukan dalam satu tahun terakhir; dan atau
 - d) Pendapatan kotor gabungan orangtua/wali dibagi jumlah anggota keluarga sebesar-besarnya Rp750.000,00 setiap bulannya;
5. Pendidikan orang tua/wali setinggi-tingginya S1 (Strata 1) atau Diploma 4.
6. Berpotensi akademik baik berdasarkan rekomendasi kepala sekolah.
7. Pendaftar difasilitasi untuk memilih salah satu diantara PTN atau PTS dengan ketentuan:
 - a) PTN dengan pilihan seleksi masuk:
 - i. Seleksi Masuk Perguruan Tinggi Negeri (SNMPTN);
 - ii. Seleksi Bersama Masuk Perguruan Tinggi Negeri (SBMPTN);
 - iii. Seleksi mandiri di 1(satu) PTN.
 - b) PTS dengan pilihan seleksi masuk di 1 (satu) PTS.

3. DATA MINING

Data mining adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan machine learning untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai

database besar. Istilah data mining memiliki hakikat sebagai disiplin ilmu yang tujuan utamanya adalah untuk menemukan, menggali, atau menambang pengetahuan dari data atau informasi yang kita miliki. Data mining, sering juga disebut sebagai Knowledge Discovery in Database (KDD). (Hadi Suyono 2013)

Data mining mulai dikenal sejak tahun 1990, ketika pekerjaan pemanfaatan data menjadi sesuatu yang penting dalam berbagai bidang, mulai dari bidang akademik, bisnis, hingga medis. Data mining dapat diterapkan pada berbagai bidang yang mempunyai sejumlah data, tetapi karena wilayah penelitian dengan sejarah yang belum lama, dan belum melewati masa remaja, maka data mining masih diperdebatkan posisi bidang pengetahuan yang memilikinya. Maka, Daryl Pregibeon menyatakan bahwa Data mining adalah campuran dari statistik, kecerdasan buatan, dan riset basis data yang masih berkembang. Terlepas dari remajanya, ternyata data mining diproyeksikan menjadi jutaan dolar di dunia industri pada tahun 2000, sedangkan pada saat yang sama ternyata data mining dipandang sebelah mata oleh sejumlah peneliti sebagai dirty word in statistic. Mereka adalah orang-orang yang tidak memandang data mining sebagai sesuatu yang menarik bagi mereka pada saat itu (Hermawati 2013).

Data mining (DM) yang juga dikenal sebagai Knowledge Discovery (Frawley et al., 1992), merupakan salah satu bidang yang berkembang pesat karena besarnya kebutuhan akan nilai tambah dari database skala besar yang makin banyak terakumulasi sejalan dengan pertumbuhan teknologi informasi. Secara umum, data mining dapat didefinisikan sebagai suatu rangkaian proses untuk menggali nilai tambah berupa ilmu pengetahuan yang selama ini tidak diketahui secara manual dari suatu kumpulan data (Hermawati 2013)

A. Pekerjaan dalam Data Mining

1) *Prediksi* :Pekerjaan ini berkaitan dengan pembuatan sebuah model yang

dapat melakukan pemetaan dari setiap himpunan variabel ke setiap targetnya, kemudian menggunakan model tersebut untuk memberikan nilai target pada himpunan baru yang diperoleh. Ada dua jenis model prediksi, yaitu klasifikasi dan regresi. Klasifikasi digunakan untuk variabel target diskrit, sedangkan regresi digunakan untuk variabel target kontinu.

2) *Analisis Pengklusteran*: Analisis kelompok melakukan pengelompokan data ke dalam sejumlah kelompok berdasarkan kesamaan karakteristik masing-masing data pada kelompok-kelompok yang ada. Data-data yang masuk dalam batas kesamaan dengan kelompoknya akan bergabung dalam kelompok tersebut, dan akan terpisah dalam kelompok yang berbeda jika keluar dari batas kesamaan kelompok.

3) *Analisis Asosiasi*: Analisis asosiasi digunakan untuk menemukan pola yang menggambarkan kekuatan hubungan fitur dalam data. Pola yang ditemukan biasanya mempersentasikan bentuk aturan implikasi atau subset fitur. Tujuannya untuk menemukan pola yang menarik dengan cara yang efisien. Tugas asosiasi dalam data mining adalah menemukan atribut yang muncul dalam satu waktu. Dalam dunia bisnis lebih umum disebut analisis keranjang belanja (Prasetyo 2014).

4) *Deteksi Anomali (anomaly detection)*: Pekerja deteksi anomali berkaitan dengan pengamatan sebuah data dari sejumlah data yang secara signifikan mempunyai karakteristik yang berbeda dari sisa data yang lain. Data-data yang karakteristiknya menyimpang (berbeda) dari data yang lain disebut sebagai outlier.

B. Knowledge Discovery from Data (KDD)

Proses KDD/ Data Mining merupakan sebuah proses mengubah data mentah menjadi suatu informasi yang berguna. Adapun proses Tahapan data mining dibagi menjadi enam bagian yaitu :

1) *Pembersihan data (data cleaning)*: sebelum proses data mining dapat dilaksanakan, perlu dilakukan proses

cleaning pada data yang menjadi fokus KDD.

2) Integrasi data (data integration): integrasi data merupakan penggabungan data dari berbagai database ke dalam satu database baru.

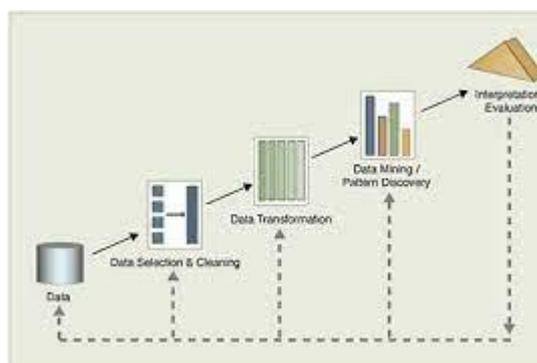
3) Seleksi Data (Data Selection): data yang ada pada database sering kali tidak semuanya dipakai, oleh karena itu hanya data yang sesuai untuk dianalisis yang akan diambil dari database.

4) Transformasi data (Data Transformation): data diubah atau digabung ke dalam format yang sesuai untuk diproses dalam data mining.

5) Proses mining: proses mining adalah sebuah proses yang paling utama pada saat metode diterapkan untuk mencari pengetahuan tersembunyi dan berharga dari data.

6) Evaluasi pola (pattern evaluation): untuk mengidentifikasi pola-pola menarik ke dalam knowledge based yang ditemukan.

7) Presentasi pengetahuan (knowledge presentation): merupakan penyajian dan visualisasi pengetahuan mengenai metode yang digunakan untuk memperoleh pengetahuan yang diperoleh pengguna.



Gbr. 1 KDD dan Data Mining

I. METODE KLASIFIKASI NAIVE BAYES CLASSIFIER

1) *Klasifikasi* : klasifikasi adalah suatu proses untuk menemukan model atau fungsi yang menguraikan atau membedakan konsep atau kelas data,

dengan tujuan untuk dapat memperkirakan kelas dari suatu objek yang labelnya tidak diketahui. Untuk mencapai tujuan tersebut, proses klasifikasi membentuk suatu model yang mampu membedakan data ke dalam kelas-kelas yang berbeda berdasarkan aturan atau fungsi tertentu. Model itu sendiri bisa berupa aturan jika-maka, pohon keputusan, atau formula matematis. (Kemenristekdikti 2015). Klasifikasi 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. Algoritma klasifikasi yang banyak digunakan secara luas, yaitu Decision/classification trees, Bayesian classifiers/ Naive Bayes classifiers, Neural networks, Analisa Statistik, Algoritma Genetika, Rough sets, k-nearest neighbor, Metode Rule Based, Memory based reasoning, dan Support vector machines (SVM).

2) *Naive Bayes Classifier*: naive Bayes adalah suatu metode klasifikasi dalam data mining dengan menggunakan metode probabilitas dan statistik sesuai dengan di kemukakan oleh ilmuwan inggris bernama Thomas Bayes. Kemudian Menurut Olson: 2008 menjelaskan bahwa Naive Bayes merupakan suatu kelas keputusan, dengan menggunakan perhitungan probabilitas matematika dengan syarat bahwa nilai keputusan adalah benar, berdasarkan informasi obyek.

3) *Bayesian classification* adalah pengklasifikasian statistik yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu class. *Bayesian classification* didasarkan pada teorema Bayes yang memiliki kemampuan klasifikasi serupa dengan decision tree dan neural network. Bayesian classification terbukti memiliki akurasi dan kecepatan yang tinggi saat diaplikasikan ke dalam database dengan data yang besar. Metode Naive Bayes merupakan metode probabilistik

pengklasifikasian sederhana berdasarkan Teorema Bayes dimana pengklasifikasian dilakukan melalui training set sejumlah data secara efisien. Nave bayes mengasumsikan bahwa nilai dari sebuah input atribut pada kelas yang diberikan tidak tergantung dengan nilai atribut yang lain. Teorema Bayes sendiri dikemukakan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya sehingga dikenal sebagai Teorema Bayes. Teorema Bayes memiliki bentuk umum sebagai berikut :

$$P(C|X) = \frac{P(X|C).P(c)}{P(X)} \quad (1)$$

Dimana :

X : Data dengan kelas yang belum diketahui

C : Hipotesis data X merupakan suatu kelas spesifik

$P(C|X)$: Probabilitas hipotesis C berdasar kondisi X (probabilitas posterior)

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

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

$P(X)$: Probabilitas X

Untuk menjelaskan teorema naïve bayes, perlu diketahui bahwa proses klasifikasi memerlukan sejumlah petunjuk untuk menentukan kelas apa yang cocok bagi sampel yang dianalisis tersebut. Karena itu, teorema bayes pada persamaan (1) disesuaikan menjadi persamaan (2) :

$$P(C|F_1 \dots F_n) = \frac{P(c)P(F_1 \dots F_n | C)}{P(F_1 \dots F_n)} \quad (2)$$

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 diatas dapat pula ditulis secara sederhana pada persamaan (3):

$$Posterior = \frac{prior \times likelihood}{evidence} \quad (3)$$

Nilai evidence selalu tetap untuk setiap kelas pada satu sampel. Nilai dari posterior tersebut nantinya akan dibandingkan dengan nilai nilai posterior kelas lainnya untuk menentukan ke kelas apa suatu sampel akan diklasifikasikan. Penjabaran lebih lanjut rumus Bayes tersebut dilakukan dengan menjabarkan ($C | F_1, \dots, F_n$) menggunakan aturan perkalian sebagai berikut:

$$\begin{aligned} P(C|F_1, \dots, F_n) &= P(C)P(F_1, \dots, F_n|C) \\ &= P(C)P(F_1|C)P(F_2, \dots, F_n|C, F_1) \\ &= P(C)P(F_1|C)P(F_2|C, f_1)P(F_3, \dots, F_n|C, F_1 F_2) \\ &= P(C)P(F_1|C)P(F_2|C, f_1)P(F_3|C, F_1 F_2)P(F_4, \dots, F_n|C, F_1 F_2 F_3) \\ &= P(C)P(F_1|C)P(F_2|C, f_1)P(F_3|C, F_1 F_2) \dots P(F_n|C, F_1 F_2 F_3, \dots, F_n - 1) \end{aligned} \quad (4)$$

Dapat dilihat bahwa hasil penjabaran tersebut menyebabkan semakin banyaknya faktor faktor syarat yang mempengaruhi nilai probabilitas, yang hampir mustahil untuk dianalisa satu persatu. Akibatnya, perhitungan tersebut menjadi sulit untuk dilakukan. Disinilah digunakan asumsi independensi yang sangat tinggi (naif), bahwa masing masing petunjuk (F_1, F_2, \dots, F_n) saling bebas (independen) satu sama lain. Dengan asumsi tersebut, maka berlaku suatu persamaan sebagai berikut:

$$\text{Untuk } P(F_i|F_j) = \frac{P(F_i \cap F_j)}{P(F_j)} = \frac{P(F_i)P(F_j)}{P(F_j)} = P(F_i) \quad (5)$$

$$P(P_i|C, F_i) = P(F_i|C) \quad (6)$$

Dari persamaan di atas dapat disimpulkan bahwa asumsi independensi naif tersebut membuat syarat peluang menjadi sederhana, sehingga perhitungan menjadi mungkin untuk dilakukan. Selanjutnya, penjabaran $P(C|F_1, \dots, F_n)$ dapat disederhanakan menjadi:

$$P(C|F_1, \dots, F_n) = P(C)P(F_1|C)P(F_2|C) \dots P(F_n|C) = P(C)\pi^n = 1 \quad P(F_i|C) \quad (7)$$

Persamaan di atas merupakan model dari teorema Naive Bayes yang selanjutnya akan digunakan dalam proses klasifikasi. Untuk klasifikasi dengan data kontinyu digunakan rumus Densitas Gauss:

$$P(X_i = x_i | Y = y_i) = \frac{(1)}{(\sqrt{2\pi}\sigma_j)} e^{-\frac{(x_i - \mu_j)^2}{(2\sigma_j^2)}} \quad (8)$$

(8)

Keterangan:

P = Peluang

X_i = Atribut ke- i

x_i = Nilai Atribut ke- i

Y = Kelas yang dicari

y_i = Sub-kelas yang dicari

μ = mean, menyatakan rata-rata dari seluruh atribut

σ = Deviasi standar, menyatakan varian dari seluruh atribut

Naive Bayes merupakan metode yang menggunakan pendekatan probabilitas untuk menghasilkan klasifikasi. Metode ini menggabungkan probabilitas term dengan probabilitas kategori untuk menentukan kemungkinan kategori

berhasil. Dari penjelasan mengenai algoritma Naive Bayes, dapat disimpulkan langkah langkah pengerjaan Naive Bayes adalah Tentukan data latih dan data uji yang ingin diklasifikasikan. Menghitung $P(C_i)$ yang merupakan probabilitas prior untuk setiap sub kelas C yang dihasilkan menggunakan persamaan:

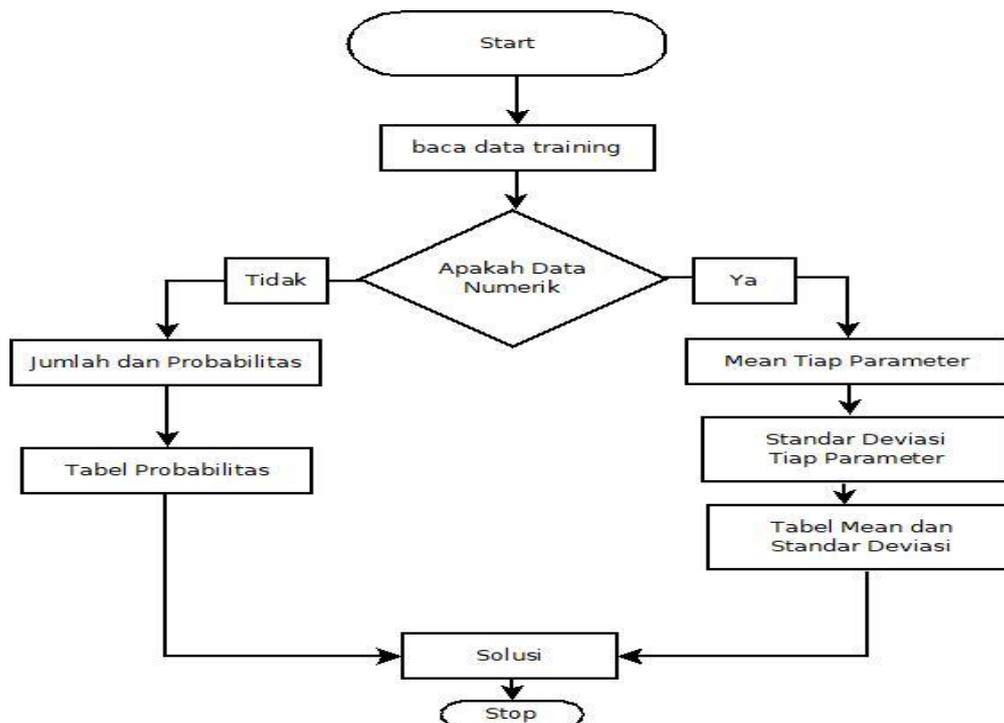
$$P(C_1) = \frac{S_i}{s} \quad (9)$$

Dimana S_i adalah jumlah data training dari kategori C_i , dan s adalah jumlah total data training. Menghitung $P(F_i|C_i)$ yang merupakan probabilitas posterior F_i dengan syarat C menggunakan persamaan (7). Apabila F_i merupakan data numerik, maka untuk menghitung $P(F_i|C_i)$ menggunakan distribusi Gaussian yang terdapat pada persamaan (8).

Memaksimalkan $P(F_i|C_i)P(C_i)$ untuk mendapatkan kelas C yang ingin diklasifikasikan dengan cara mengkalikan $P(F|C_i)$ dan $P(C_i)$ untuk semua kemungkinan klasifikasi:

$$P(C_i)\pi^n = 1 \quad P(F_i|C_i) \quad (10)$$

Dengan kata lain, hasil yang ditetapkan ke dalam kelas C_i adalah yang mempunyai $P(F_i|C_i)P(C_i)$ maksimum.



Gbr. 1 Skema Naive Bayes

σ = standar deviasi

Skema pada gambar 2 menerangkan proses naivet bayes :

- 1) Baca data training.
- 2) Hitung jumlah dan probabilitas, namun apabila data numerik maka

Cari nilai mean dan standar deviasi dari masing-masing parameter yang merupakan data numerik. Adapun persamaan yang digunakan untuk menghitung nilai rata rata (mean) dapat dilihat sebagai berikut:

$$\mu = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n} \quad (11)$$

dimana

μ = rata-rata hitung (mean)
 x_i = nilai sampel ke-i n = jumlah sampel,
 Adapun persamaan untuk menghitung standar deviasi adalah sebagai berikut:

$$\sigma = \sqrt{\frac{\sum_{i=1}^n (x_i - \mu)^2}{n - 1}} \quad (12)$$

dimana

x_i = nilai x ke-i

μ = rata-rata hitung, n = jumlah sampel

- a. Jika data berbentuk kategorikal cari nilai probabilitas dengan cara menghitung jumlah data yang sesuai dari kategori yang sama dibagi dengan jumlah data pada kategori tersebut. Adapun persamaan untuk menghitung nilai probabilitas bagi data kategorikal adalah sebagai berikut:

$$P(X_i = x_i | Y = y_1) = \frac{(\sum \text{nilai data pada kelas } X_i)}{\sum \text{data kelas } X_i} \quad (13)$$

P= Peluang

X_i = nilai atribut

x_i = atribut ke i

Y = Kelas yang di cari

y_1 = Sub kelas Y yang dicari

Mendapatkan nilai dalam tabel mean, standard deviasi dan probabilitas Penghitungan probabilitas data numerik menggunakan densitas Gauss pada

persamaan 8.

Untuk mengevaluasi performa dari model yang dibangun, perlu dilakukan pengukuran performa, yaitu pengukuran akurasi (accuracy) atau tingkat kesalahan (error rate). Jika f_{ij} menotasikan jumlah record dari kelas I yang berada di kelas j pada saat pengujian, maka pengukuran akurasi (accuracy) dapat ditulis dengan persamaan sebagai berikut

$$\begin{aligned} \text{Accuracy} &= \frac{\text{jumlah prediksi yg benar}}{\text{Jumlah prediksi keseluruhan}} \times 100\% \\ &= \frac{f_{11} + f_{00}}{f_{11} + f_{10} + f_{01} + f_{00}} \quad (14) \end{aligned}$$

$$\begin{aligned} \text{Error} &= \frac{\text{jumlah prediksi yg salah}}{\text{Jumlah prediksi keseluruhan}} \times 100\% \\ &= \frac{f_{01} + f_{10}}{f_{11} + f_{10} + f_{01} + f_{00}} \quad (15) \end{aligned}$$

II. METODOLOGI PENELITIAN

Penelitian ini akan dilakukan di Universitas Negeri Medan dan waktu yang dibutuhkan untuk penelitian ini kurang lebih selama dua bulan.

Jenis penelitian ini adalah Studi Kasus jurusan matematika di Universitas Negeri Medan. Data yang digunakan merupakan data sekunder yang diperoleh dari Kantor BAAK Universitas Negeri Medan, berupa data mahasiswa S1 Angkatan 2019 Universitas Negeri Medan. Data yang diambil antara lain pekerjaan orang tua, penghasilan orang tua, jumlah tanggungan, daya listrik (watt), dan nilai ujian nasional. Prosedur penelitian ini adalah sebagai berikut:

1. Mengumpulkan data yang diperlukan dalam penelitian yaitu data mahasiswa S1 Angkatan 2019 jurusan matematika di Universitas Negeri Medan.
2. Data Cleaning adalah proses

membersihkan data dari data noise dan tidak konsisten.

3. Data Integration adalah proses untuk menggabungkan data dari beberapa sumber yang berbeda.
4. Membaca data training
5. Menghitung Jumlah dan Probabilitas, namun apabila data numerik maka
 - a. Mencari nilai mean dan standar deviasi dari masing-masing parameter yang merupakan data numerik.
 - b. Jika data berbentuk numerik mencari nilai probabilitasnya dengan cara menghitung jumlah data yang sesuai dari kategori yang sama dibagi dengan jumlah data pada numerik tersebut
6. Mendapatkan nilai dalam tabel mean, standard deviasi dan probabilitas penghitungan probabilitas data numerik menggunakan densitas Gauss
7. Solusi kemudian di hasilkan
8. Menghitung akurasi sistem

III. PEMBAHASAN

Naïve Bayes Classifier (NBC) merupakan sebuah pengklasifikasian probabilitas sederhana yang mengaplikasikan Teorema Bayes dengan asumsi ketidaktergantungan (independent) yang tinggi. Keuntungan penggunaan NBC adalah metode ini hanya membutuhkan jumlah data pelatihan (training data) yang kecil untuk menentukan estimasi parameter yang diperlukan dalam proses pengklasifikasian. Adapun beberapa tahapan dari proses prosedur dari Klasifikasi Naïve Bayes Classifier (NBC) yaitu :

i. Pengumpulan Data

Penelitian yang dilakukan merupakan jenis penelitian studi kasus mengenai penentuan penerima beasiswa di Universitas Negeri Medan dengan menggunakan metode Naïve Bayes Classifier. Data yang digunakan diperoleh dari Kantor BAAK (Biro

Administrasi Akademik Kemahasiswaan) berupa data mahasiswa jurusan matematika angkatan 2019 dengan jumlah data 318 mahasiswa, data tersebut merupakan data sekunder dan data yang didalamnya terdapat berbagai kriteria

data yaitu : Nilai Rata-rata UN, Jumlah Anak, Pekerjaan Orangtua, Penghasilan orangtua dan IPK (Indeks Prestasi Kumulatif). Data yang diperoleh seperti pada tabel.1

TABEL I
Data Mahasiswa Jurusan Matematika Angkatan 2019
Di Universitas Negeri Medan

N o	RATA UAN	JUMLAH ANAK	KERJA AYAH	PENGHASILAN ORANG TUA	LISTRİK (WATT)	IPK
1	72.75	7	LAINNYA	3,000,000	450	3,31
2	72.62	3	LAINNYA	5,333,000	900	0
3	71.63	4	PEG. BUMN	2,667,918	450	3,09
4	69.75	4	LAINNYA	2,500,000	900	3,28
5	67.75	4	PNS PENDIDIK	2,860,000	450	3,18
.
.
.
384	69.2	4	LAINNYA	6,151,100	450	2,95
385	71.66	2	PETANI	1,000,000	450	3,28
386	62.77	5	PETANI	3,000,000	450	2,58
387	51.38	3	PETANI	60,000	450	3,28
388	73	1	PETANI	5,000,000	450	3

2) Data Cleaning

Data cleaning adalah proses untuk mengisi missing value atau isian yang hilang atau kurang tepat, mengoreksi data yang tidak konsisten, dan mendeteksi redudansi data yaitu penumpukan data atau duplikasi data. Peneliti melakukan proses data cleaning dengan memlter setiap kriteria dan mencari data yang tidak digunakan dalam data set. Adapun kriteria yang harus dicleaning yaitu 70 data mahasiswa

yang memperoleh beasiswa bidik misi dari total data keseluruhannya sebanyak 388, sehingga data berkurang menjadi 318 data yang akan digunakan dalam proses mining. Data mahasiswa yang memperoleh beasiswa bidikmisi tersebut tidak di masukan ke dalam kriteria di karenakan mahasiswa yang telah menerima beasiswa bidik misi tidak boleh menerima beasiswa lainnya.

TABEL III
DATA MAHASISWA JURUSAN MATEMATIKA ANGKATAN 2019 DI
UNIVERSITAS NEGERI MEDAN

N	RATA	JUMLAH	KERJA	PENGHASILA	LISTRİK	IPK
1	72.75	7	LAINNYA	3,000,000	450	3,31
2	72.62	3	LAINNYA	5,333,000	900	0
3	71.63	4	PEG. BUMN	2,667,918	450	3,09
4	69.75	4	LAINNYA	2,500,000	900	3,28
5	67.75	4	PNS PENDIDIK	2,860,000	450	3,18

31	1.92	4	LAINNYA	6,151,100	450	2,95
31	1.71	2	PETANI	1,000,000	450	3,28
31	1.62	5	PETANI	3,000,000	450	2,58
31	1.25	3	PETANI	60,000	450	3,28
31	0	1	PETANI	5,000,000	450	3

3) Data Transformation

Data Transformation, juga dikenal sebagai data consolidation. Pada tahap ini, dimana data-data yang telah terpilih, ditransformasikan kedalam bentuk-bentuk yang cocok untuk prosedur penggalian (meaning proedure) dengan cara melakukan normalisasi dan agregasi data. Dalam proses transformation data pada kriteria rata- rata nilai UN, jumlah anak, kerja ayah, penghasilan orangtua, listrik, IPK.

Atribut yang dibutuhkan dalam penelitian ini telah diperoleh, tetapi isi data atribut tersebut masih belum memenuhi kriteria yang cocok untuk dilanjutkan ke

dalam proses Klasifikasi Naive Bayes. Dalam proses klasifikasi Naive Bayes data yang bisa diolah hanya data yang berupa angka sehingga data tersebut harus diubah ke dalam bentuk text. Langkah pertama yang dilakukan adalah mencari frekuensi terbesar sampai terkecil. Pada atribut nilai UN peneliti memperoleh data tinggi dan layak sebanyak 197 mahasiswa, tidak layak dan tinggi sebanyak 50, sedangkan atribut nilai UN rendah dan layak sebanyak 53, dan rendah tidak layak sebanyak 68 mahasiswa. Tabel rata-rata nilai UN yang sudah di transformasikan terdapat pada tabel III.

TABEL IIIII
NILAI RATA-RATA UN(UJIAN NASIONAL)

Atribut	Layak	Tidak Layak
Tinggi	197	50
Rendah	53	68

Pada tabel Jumlah anak, Pekerjaan orangtua, Penghasilan orangtua, Pemakaian listrik dan IPK mahasiswa berisikan 318 data. Setelah di analisa terdapat data angka yaitu Jumlah anak dan

diberikan inisial pada setiap data berdasarkan jumlah anak. Hasil data transformation jumlah anak dapat dilihat di Tabel IV.

TABEL IV
JUMLAH ANAK

Jumlah anak	Atribu dan Insial			
	Tinggi dan Layak	Renda dan Layak	Tinggi dan Tidak Layak	Rendah dan Tidak Layak
1 sampai 7		247		68
8 sampai 14	3		0	

Setelah di analisa selain data Jumlah anak, terdapat format yang tidak sesuai juga pada data Pekerjaan Orangtua, sehingga diubah dalam format Atribut.

Pekerjaan orangtua di klasifikasi menjadi 12 kriteria dari 318 data berdasarkan fropesi pekerjaan orangtua mahasiswa.

Hasil data transformation pekerjaan orangtua dapat dilihat pada tabel V.

TABEL V JUMLAH ANAK

Pekerjaan Orangtua	Atribut	
	Layak	Tidak Layak
Buruh	11	4
Nelayan	2	0
Pengawai BUMN	7	2
Pengawai BUMS	8	1
Pensiun	10	2
Petani	40	12
PNS Nondik	9	3
PNS Pendidk	30	4
TNI/POLRI	8	3
Upah Harian	9	2
Wiraswasta	32	5
Lainnya	84	30

Selain data Jumlah anak dan Pekerjaan dilakukan penginsialan seperti pada tabel orangtua terdapat juga format penghasilan VI. orangtua yang tidak sesuai, sehingga

TABEL VI
DATA PENGHASILAN ORANGTUA

Atribut&Insial	Penghasilan Orangtua		
	0 - 5.900.000	6.000.000-10.900.000	7.000.000-15.900.000
Tinggi&Layak			3
Sedang&Layak		57	
Rendah&Layak	190		
Tinggi&Tidak Layak			2
Sedang&Tidak Layak		18	
Rendah&Tidak Layak	48		

Selain Penghasilan orangtua terdapat juga seperti pada tabel 7. format Jumlah Pemakaian Listrik, sehingga dilakukan penginsialan

TABEL VI
DATA PEMAKAIAN LISTRIK(WATT)

Atribut & Inisial	Listrik(Watt)		
	450	900	130
Tinggi&Layak			20
Sedang&Layak		36	
Rendah&Layak	194		
Tinggi&Tidak			9
Sedang& Tidak		6	
Rendah&Tidak	47		

Selain Pemakaian listrik, terdapat juga IPK yang format nya dilakukan penginisialan seperti pada tabel VIII

TABEL VIII
DATA INDEK PRESTASI KUMULATIF

Atribut & Inisial	IPK	
	0-	3.00-
Tinggi&Layak	248	
Rendah&Layak		2
Tinggi&Tidak	55	
Rendah&Tidak		13

4) Proses Perhitungan Naïve Bayes Classifier

Tahap ini adalah tahap dimana akan dilakukan proses perhitungan data penentuan beasiswa mahasiswa jurusan matematika angkatan 2019 di Universitas Negeri Medan dengan menggunakan metode Naïve Bayes.

Tahapan proses perhitungannya sebagai berikut:

- 1) Menghitung Jumlah Class/Label

$$P(B=Layak) = \frac{250}{318} = 0,78$$

$$P(B=Tidak Layak) = \frac{250}{318} = 0,21$$

2) Menghitung jumlah kasus yang sama dengan class yang sama.

$$P(\text{Rata-rata UN Tinggi \& Layak}) = \frac{197}{250} = 0,788$$

$$P(\text{Rata-rata UN Tinggi \& Tidak Layak}) = \frac{50}{68} = 0,73$$

$$P(\text{Rata-rata UN Tidak Layak}) = \frac{18}{68} = 0,265$$

$$P(\text{Rata-rata UN Layak}) = \frac{197}{250} = 0,788$$

$$P(\text{Rata-rata UN Tidak Layak}) = \frac{18}{68} = 0,265$$

$$P(\text{Rata-rata UN Layak}) = \frac{197}{250} = 0,788$$

$$P(\text{Rata-rata UN Tidak Layak}) = \frac{18}{68} = 0,265$$

$$P(\text{Rata-rata UN Layak}) = \frac{197}{250} = 0,788$$

$$P(\text{Rata-rata UN Tidak Layak}) = \frac{18}{68} = 0,265$$

3) Mengalikan semua hasil variabel setiap klasifikasi

$$P(\text{Rata-rata UN Layak}) * P(\text{Jumlah Anak Layak}) * P(\text{Pekerjaan Orangtua Layak}) * P(\text{Penghasilan Orangtua Layak}) * P(\text{Pemakaian Listrik(Watt) Layak}) * P(\text{Indeks Prestasi Kumulatif Layak})$$

$$= \frac{197}{250} * \frac{53}{250} * \frac{3}{250} * \frac{247}{250} * \frac{11}{250} * \frac{2}{250} * \frac{7}{250} * \frac{8}{250} * \frac{10}{250} * \frac{40}{250} * \frac{9}{250} * \frac{30}{250} * \frac{8}{250} * \frac{9}{250} * \frac{3}{250} * \frac{84}{250} * \frac{3}{250} * \frac{57}{250} * \frac{190}{250} * \frac{20}{250} * \frac{36}{250} * \frac{194}{250} * \frac{248}{250} * \frac{2}{250} = 5,04919$$

$P(\text{Rata-rata UN Tidak Layak}) * P(\text{Jumlah Anak Tidak Layak}) * P(\text{Pekerjaan Orangtua Tidak Layak}) * P(\text{Penghasilan Orangtua Tidak Layak}) * P(\text{Pemakaian Listrik(Watt) Tidak Layak}) * P(\text{Indeks Prestasi Kumulatif Tidak Layak})$

$$= \frac{18}{68} * \frac{18}{68} * \frac{0}{68} * \frac{68}{68} * \frac{4}{68} * \frac{0}{68} * \frac{2}{68} * \frac{1}{68} * \frac{2}{68} * \frac{12}{68} * \frac{3}{68} * \frac{4}{68} * \frac{3}{68} * \frac{2}{68} * \frac{5}{68} * \frac{30}{68} * \frac{2}{68} * \frac{18}{68} * \frac{48}{68} * \frac{9}{68} * \frac{6}{68} * \frac{47}{68} * \frac{55}{68} * \frac{13}{68} = 0$$

4) Membandingkan setiap hasil class dalam klasifikasi

Dari hasil diatas, terlihat bahwa nilai probabilitas tertinggi ada pada kelas (P_{Layak}) dengan nilai 5,04919 sehingga dapat disimpulkan bahwa penentuan penerima beasiswa dikategorikan layak sebanyak 250 mahasiswa dan tidaklayak sebanyak 68 mahasiswa dari seluruh 318 mahasiswa.

1V. KESIMPULAN

Dari hasil penelitian ini, dapat diambil kesimpulan sebagai berikut

1. Metode Naive Bayes Classification digunakan untuk mengklasifikasikan Penerima beasiswa menghasilkan akurasi yang baik, dengan nilai kelayakannya 5,04919.
2. Pengujian pada perbandingan data training dan data testing sebesar 80:20 menghasilkan akurasi tertinggi dengan 79%, dan penentuan penerima beasiswa dikategorikan layak sebanyak 250 mahasiswa dan tidaklayak sebanyak 68 mahasiswa dari seluruh 318 mahasiswa.

Kemaha- siswaan, ANDI, Jakarta, 2015

- [6] Lahinta, *Konsep Rancangan Sistem Pendukung Keputusan Penentuan Kandidat Penerima Beasiswa Studi Kasus pada TPSDM Propinsi Gorontalo*, Vol. 1 of 7, Universitas Gadjah Mada, Yogyakarta, 2009.
- [7] Prasetyo, E., *Data Mining: Konsep dan Aplikasi Menggunakan Matlab*, ANDI, Yogyakarta, 2014.

REFERENSI

- [1] H Suyono, M Ridwan, d. M. S., *Penerapan Data Mining Untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier*, Vol. 1 of 7, EECCIS, Yogyakarta. 2013.
- [2] Hendri, *Penerapan Metode K-Nearest Neighbor pada aplikasi penentu penerima beasiswa mahasiswa di STMIK Sinar Nusantara Surakarta*, STMIK Sinar Nusantara Surakarta, Yogyakarta. 2014.
- [3] Hermawati, F., *Data Mining*, Vol. 1, ANDI, Yogyakarta, 2013.
- [4] Junaidi (2017): *Pemilihan Penerima Beasiswa Menggunakan Metode Profile Matching*, Vol. 2 of 19, Paradigma, Yogyakarta, 2017.
- [5] Kemenristekdikti, *Pedoman Umum Beasiswa dan Bantuan Biaya Pendidikan Bidikmisi*, Direktorat Jenderal Pembelajaran dan