IMPLEMENTASI METODE NAÏVE BAYES UNTUK MENENTUKAN PERSETUJUAN PEMBERIAN BEASISWA PENUH PADA PENERIMAAN MAHASISWA BARU DI INSTITUSI PENDIDIKAN X

Wilianti Aliman

Sekolah Tinggi Manajemen Informatika & Komputer LIKMI
Jl. Ir. H. Juanda No. 96 Bandung

E-mail : wilty@likmi.ac.id

ABSTRAK

Beasiswa merupakan salah satu strategi yang dapat digunakan untuk meningkatkan jumlah calon mahasiswa yang mendaftar di sebuah institusi Pendidikan terutama Institusi Pendidikan Swasta. Beasiswa yang diberikan juga harus disesuaikan agar tidak merugikan bisnis dari Institusi Pendidikan tersebut, sehingga perlu menganalisa karakteristik calon penerima beasiswa. Institusi Pendidikan harus dengan cermat menentukan karakteristik calon mahasiswa yang berpotensi untuk diberikan beasiswa.

Penelitian ini memanfaatkan proses Data Mining untuk menemukan dan mempelajari pola perilaku yang ada dari data masa lalu. Pengetahuan akan profil pelanggan yang berhasil menjadi mahasiswa dengan yang kemungkinan besar tidak akan menjadi mahasiswa ini dapat digunakan untuk mengukur perilaku calon mahasiswa yang mendaftar di Institusi Pendidikan X dengan melihat demografi dari calon mahasiswa, waktu pendaftaran, dan juga besaran beasiswa yang telah diberikan dari pengalaman yang sudah terjadi khususnya data pada tahun 2016-2020.

Penelitian menggunakan metode Naïve Bayes ini menghasilkan tingkat akurasi 83,3%, menurut penelitian yang dilakukan *probability* keberhasilan menjadi mahasiswa (calon mahasiswa menuntaskan proses pendaftran hingga menjadi mahasiswa) lebih tinggi apabila pemberian beasiswa dilakukan melalui ujian atau prestasi dari calon mahasiswa.

ABSTRACT

Scholarships are a strategy that can be used to increase the number of prospective students who enroll in an educational institution, especially private educational institutions. Scholarships given must also be adjusted so as not to harm the business of the educational institution, so it is necessary to analyze the characteristics of prospective scholarship recipients. Educational institutions must carefully determine the characteristics of prospective students who have the potential to be awarded a scholarship.

This research utilizes the Data Mining process to find and study existing behavior patterns from past data. Knowledge of the profiles of customers who have successfully become students and those who are unlikely to become students can be used to measure the behavior of prospective students who enroll in Educational Institution X by looking at the demographics of prospective students, the time of registration, and also the amount of scholarships that have been awarded from past experience. This has happened, especially the data in 2016-2020.

Research using the Naïve Bayes method produces an average accuracy rate of 83,3%, according to research conducted the probability of success as a student (prospective students complete the registration process to become students) is higher if the scholarship is awarded through examinations or achievements of prospective students.

Kata kunci : *Naïve Bayes, Data Mining, Classification*

1. PENDAHULUAN

Penerimaan Mahasiswa Baru merupakan salah satu proses bisnis yang terdapat pada institusi Pendidikan. Proses ini selalu dilakukan setiap tahunnya oleh setiap Institusi Pendidikan untuk mempertahankan eksistensi dari Institusi Pendidikan tersebut. Penelitian yang dilakukan oleh Hananto menjelaskan bahwa peran Data Mining pada Institusi pendidikan sangatlah penting karena dengan adanya pengolahan data sebuah Institusi pendidikan mendapatkan pengetahuan baru mengenai pelanggannya yang dalam hal ini adalah calon mahasiswa [3]. Ketika mengenal setiap mahasiswa maupun calon mahasiswa, setiap marketing dapat mengembangkan metode untuk melakukan promosi dan interaksi yang lebih sesuai target atau menyesuaikan dengan perilaku calon mahasiswa. Hasil pengolahan data juga dapat digunakan untuk menentukan produk/ program yang efektif untuk dilakukan terlebih dahulu dan juga dapat menentukan pelanggan yang memiliki potensi paling menguntungkan untuk dipertahankan.

Masalah yang sering terjadi banyak calon mahasiswa sering membatalkan pendaftaran walau calon mahasiswa tersebut sudah diberikan beasiswa hingga seluruh biaya perkuliahan. Beasiswa yang diberikan juga harus disesuaikan agar tidak merugikan bisnis dari Institusi Pendidikan tersebut, sehingga perlu menganalisa karakteristik calon penerima beasiswa. Institusi Pendidikan harus dengan cermat menentukan karakteristik calon mahasiswa yang berpotensi untuk diberikan beasiswa. Pada penelitian yang dilakukan Salim menjelaskan bahwa adanya hubungan mengenai demografi calon mahasiswa dengan karakteristik sebuah Kampus [8]. Hal tersebut memberi pengaruh pada proses pendaftaran dan proses penerimaan mahasiswa baru. Analisa mengenai karakteristik calon mahasiswa juga dapat membantu Institusi Pendidikan tersebut menemukan strategi baru yang lebih cocok untuk mencapai target yang diharapkan dan juga sebagai bahan evaluasi kedepannya.

Penelitian ini memanfaatkan proses Data Mining dengan metode Naïve Bayes untuk menemukan dan mempelajari pola perilaku yang ada dari data masa lalu. Metode Naïve Bayes memiliki performa yang baik dalam melakukan *Classification* yaitu hingga 70% lebih baik dibanding metode lainnya seperti K-NN atau Decission Tree [5]. Metode Naïve Bayes memanfaatkan perhitungan probabilitas dan statistik sehingga dapat dimanfaatkan juga untuk melihat probabilitas dari sebuah kelompok perilaku calon mahasiswa untuk benar-benar menjadi mahasiswa di Institusi Pendidikan. Menurut Yuda Septian Nugroho, Naïve Bayes merupakan algoritma yang mampu memberikan hasil dengan bukti hasil akurasi dan kecepatan yang tinggi ketika diaplikasikan pada database [4]. Kelebihan dari metode Naïve Bayes tersebut sesuai dengan tujuan dari penelitian ini yaitu menganalisis probabilitas dari kelompok calon mahasiswa yang berhasil menjadi mahasiswa dengan yang kemungkinan besar tidak akan menjadi mahasiswa.

Data Set yang akan digunakan pada penelitian ini yaitu dengan menggunakan data demografi dari calon mahasiswa, waktu pendaftaran, dan juga besaran beasiswa yang telah diberikan dari pengalaman yang sudah terjadi khususnya data pada tahun 2016-2020. Pola-pola yang ditemukan tersebut akan dipisahkan menjadi beberapa segmentasi calon mahasiswa. Segmentasi atau *classification* pada Data Mining dilakukan dengan cara mengelompokan data yang sama atau mirip. Segmentasi bertujuan untuk menentukan kelompok-kelompok margin pelanggan berdasarkan kebutuhan dan perilakunya.

2. RUMUSAN MASALAH DAN TUJUAN

Rumusan Masalah pada penelitian ini adalah sebagai berikut:

a. Profil calon mahasiswa seperti apa yang berhasil menjadi mahasiswa Institusi X?
b. Profil calon mahasiswa seperti apa yang sesuai untuk diberikan Beasiswa penuh dengan Beasiswa yang hanya Potongan Uang Pangkal?

Tujuan pada penelitian ini adalah sebagai berikut:

a. Mengetahui informasi profil setiap calon mahasiswa yang berhasil menjadi mahasiswa
b. Menentukan kriteria untuk mempermudah pengambilan keputusan untuk pemberian Beasiswa pada calon mahasiswa.

3. LANDASAN TEORI

3.1. *Knowledge Discovery in Databases (KDD)*

KDD merupakan salah satu alat bantu proses Data Mining yang sudah banyak digunakan sejak dulu dan bahkan banyak digunakan pada penelitian-penelitian baru-baru ini [2]. KDD adalah proses kompleks terkait penemuan pola dalam sekumpulan data yang ada [7]. KDD mampu melakukan pemetaan data yang tidak terstruktur menjadi data yang terstruktur dan rapih. Usama Fayyad dkk. memberikan sebuah pendapat mengenai KDD sebagai berikut: *"The basic problem addressed by the KDD process is one of mapping low-level data into other forms that might be more compact, more abstract, or more useful."* [2]

Alasan penggunaan KDD sebagai alat bantu adalah dikarenakan konsep dan tujuan hasil yang sama. Tujuan dari KDD adalah melakukan pengolahan data sehingga didapati satu informasi baru untuk menentukan strategi atau keputusan untuk dunia marketing. KDD juga memiliki konsep yang sama, fungsi dari KDD adalah mengolah data dari sebuah data yang jumlahnya banyak sehingga ditemukan pola-pola yang selanjutnya dapat dilakukan segmentasi sehingga menghasilkan informasi baru. Informasi yang dihasilkan tersebut dapat dikaitkan atau diimplematasikan pada banyak bidang, salah satunya dalam pemasaran. Pada Marketing, KDD dapat membantu pengolahan data pelanggan untuk mengidentifikasi kelompok pelanggan yang berbeda dan meramalkan perilaku mereka [2]

KDD berfokus pada keseluruhan proses penemuan pengetahuan dari data, termasuk bagaimana data disimpan dan diakses, bagaimana algoritma dapat ditingkatkan ke aplikasi kumpulan data besar-besaran. Pada gambar 2.3 menjelaskan setiap tahapan pada KDD, proses-proses tersebut dijelaskan oleh oliveira sebagai berikut [7]:

1. *Data Selection*
   Tahap ini merupakan proses pemilihan data dan pemahaman mengenai penngaplikasian domain. Pengaplikasian domain bertujuan untuk memberikan konteks proyek dalam operasi perusahaan, dengan memahami bahasa bisnis dan menentukan tujuan proyek. Sedangkan pemilihan data bertujuan untuk memfokuskan analisis pada subset variabel atau sampel data, di mana penemuan harus dilakukan. Pada tahap ini, perlu untuk mengevaluasi subset minimum data yang akan dipilih, atribut yang relevan dan periode waktu yang tepat untuk dipertimbangkan.
2. *Data Processing*
   Tahap ini mencakup operasi dasar, seperti: menghilangkan noise atau outlier, mengumpulkan informasi yang diperlukan untuk memodelkan atau memperhitungkan kebisingan, memutuskan strategi untuk menangani atribut data yang hilang, dan menghitung informasi urutan waktu dan perubahan yang diketahui. Tahap ini juga mencakup masalah yang berkaitan dengan sistem

manajemen basis data, seperti tipe data, skema, dan pemetaan nilai yang hilang dan tidak diketahui.

3. *Data Transformation*
   Tahap ini merupakan tahap pemrosesan menjadi data yang memiliki format yang sesuai untuk dilakukan *Data Mining*. Transformasi yang paling umum adalah: normalisasi data, agregasi data dan diskritisasi data. Normalisasi dilakukan pada setiap nilai dikurangi rata-rata dan dibagi dengan standar deviasi. Normalisasi dilakukan Ketika Data berasal dari beberapa tabel, karena beberapa algoritma hanya dapat menangani data kuantitatif dan beberapa lagi menangani data kualitatif. Oleh karena itu, mungkin perlu untuk mendiskritisasi data, yaitu memetakan data kualitatif ke data kuantitatif, atau memetakan data kuantitatif ke data kualitatif.
4. *Data Mining*
   Tahap ini terdiri dari menemukan pola dalam dataset yang sebelumnya disiapkan. Pada tahap ini juga dilakukan pemilihan algoritma yang tepat sesuai dengan kebutuhan, kemudian diterapkan pada data yang relevan, untuk menemukan hubungan implisit atau pola menarik lainnya.
5. *Interpreter / Evaluation*
   Tahap ini terdiri dari menafsirkan pola yang ditemukan dan mengevaluasi kegunaan dan kepentingannya sehubungan dengan pengaplikasian domain.

KDD merupakan suatu lingkung yang lebih besar untuk menemukan sebuah pengetahuan baru dari data yang ada yang memanfaatkan Data Mining untuk mengekstrak polanya. Singkatnya KDD merupakan tahapan atau langkah-langkah yang digunakan untuk mengolah dan menjelajahi penemuan dari data, sedangkan Data Mining merupakan pengaplikasian sebuah algoritma untuk menghasilkan pola-pola untuk KDD menemukan penemuan-penemuan dari data tersebut.

3.2. *Data Mining*

Data Mining merupakan tipe analisis yang lebih tinggi lagi, Data Mining menganalisis data dan memberikan informasi dengan pendekatan data seperti statistik deskriptif (frekuensi, rata-rata, median, mode, varian, standar deviasi), reduksi data, analisis statistik bivariat (tabulasi silang, korelasi), analisis statistik multivariat (regresi berganda, analisis faktor, analisis diskriminan, analisis klaster, penskalaan multidimensi, analisis konjoin), pohon keputusan dan jaringan saraf, dan visualisasi data [1]. Data Mining dapat diimplementasikan untuk keperluan pemasaran terutama untuk mengenal karakter dari calon pelanggan. Data Mining merupakan proses yang menggunakan teknik stastistik, matematika, kecerdasan buatan, dan Machine learning untuk melakukan pengolahan data dan menentukan sebuah informasi baru yang berguna dari sebuah kumpulan data yang besar [9].

Data Mining digunakan untuk pemasaran memproses data yang menghasilkan sejumlah informasi, diantaranya adalah menghitung jumlah pembelian, menentukan target pasar untuk produk tertentu, memprediksi jumlah customer yang akan diberikan pelayanan lebih pada tahun mendatang, memprediksi jumlah customer yang akan kehilangan loyalitasnya, mendeskripsikan karakteristik pelanggan yang memberikan kita keuntungan lebih, mendeskripsikan perilaku konsumen yang digambarkan secara multidimensi pada segmentasi kebutuhan pelanggan, memberikan gambaran umum tentang kebutuhan-kebutuhan pelanggan untuk setiap segmen kebutuhan pelanggan, dan juga memberikan informasi panjang umur dari nilai yang diberikan oleh setiap pelanggan [1].

Data Mining memiliki beberapa cakupan-cakupan teknik mengolah data sesuai dengan fungsinya masing-masing, diantaranya yaitu [7]:

1. *Association*
   Asosiasi bertujuan untuk menentukan hubungan antara beberapa atribut dalam database. Asosiasi berfokus pada memperoleh korelasi multi-atribut, dukungan yang memuaskan, dan batas kepercayaan.
2. *Classification*
   Klasifikasi bertujuan untuk menentukan peta setiap item data ke dalam sebuah kategori yang telah ditentukan. Sebagai contoh, klasifikasi dapat digunakan untuk mengidentifikasi resiko pemohon pinjaman.
3. *Clustering*
   Clustering memiliki kemiripan dengan klasifikasi, perbedaanya hanya pada kategori yang ditentukan. Pada klasifikasi, kategori tersebut telah ditentukan namun pada Clustering tidak. Clustering menemukan kelompok alami dari item data, berdasarkan kesamaan metrik atau model kepadatan probabilitas
4. *Forecasting*
   Forecasting bertujuan memperkirakan nilai atribut tertentu di masa mendatang, berdasarkan pola rekaman. Pada forecasting, atribut yang diukur untuk setiap item untuk memprediksi perilaku di masa mendatang.
5. *Regression*
   Regresi bertujuan untuk memetakan setiap item data ke dalam kategori/ variable nilai prediksi.
6. *Sequence discovery*
   Sequence discovery bertujuan mengidentifikasi hubungan antar item dari waktu ke waktu Ini pada dasarnya dapat dianggap sebagai penemuan asosiasi atas basis data sementara. Misalnya, analisis urutan dapat dikembangkan untuk menentukan, jika pelanggan telah mendaftar untuk rencana A, maka rencana selanjutnya yang akan diambil oleh pelanggan dan dalam satu kerangka waktu diidentifikasi dalam *Sequence Discovery*.
7. *Visualization*
   Visualisasi digunakan untuk menyajikan data sehingga pengguna dapat melihat pola yang lebih kompleks, Visualisasi digunakan bersama dengan model *Data Mining* lainnya untuk memberikan pemahaman yang lebih jelas tentang pola atau hubungan yang ditemukan.

Data Mining memiliki dua tipe teknik mempelajari pola, yaitu: *Suppervised* dan *Unsupervised* [7]. Suppervised merupakan teknik pengelompokan data yang memiliki syarat bahwa dataset berisi ciri-ciri dan kebiasaan yang akan diprediksikan. Sebagai contoh, model yang diawasi dapat dilatih untuk mengidentifikasi pola-pola yang memungkinkan untuk mengklasifikasikan klien bank sebagai calon gagal bayar atau yang tidak membayar. Sedangkan *Unsuppervised* tidak memerlukan syarat seperti *Suppervised* karena *Unsuppervised* dapat dilatih untuk mengelompokkan pelanggan ke dalam kelompok yang tidak dikenal yang serupa.

Ketika melakukan *Data Mining* menggunakan metode yang dipilih ada kalanya perlu dilakukan untuk mengukur akurasi dari perhitungan yang dilakukan oleh metode tersebut. Pada bagian evaluasi untuk mengukur akurasi sebuah metode dan mengukur tingkat eror sebuah metode, maka dimanfaatkan perhitungan dari *Confusion Matrix* khususnya pada klasifikasi. Gambaran untuk *Confusion Matrix* dapat dilihat seperti pada Tabel 1.

Tabel 1
Gambaran *Confusion Matrix*

Keterangan		**Nilai yang sebenarnya**	
		True	False
Nilai Prediksi	True	TP	FP
	False	FN	TN

Hasil dari *Confusion Matrix* terdiri dari 4 istilah yaitu:

1. *True Positive* (TP)
   TP merupakan hasil prediksi sesuai dengan nilai yang sebenarnya.
2. *True Negative* (TN)
   TN merupakan hasil prediksi salah dan nilai sebenarnya juga salah
3. *False Positive* (FP)
   FP merupakan hasil prediksi benar dan nilai sebenarnya salah
4. *False Negative* (FN)
   FN merupakan hasil prediksi salah dan nilai sebenarnya benar

Berdasarkan pembahasan dari beberapa ahli dapat ditarik kesimpulan bahwa *Data Mining* dapat menjadi satu alat untuk melakukan segmentasi pelanggan terutama pada Institusi Pendidikan, karena menggunakan data yang tidak terstruktur untuk menemukan pola-pola sehingga dapat menjadi sebuah data terstruktur untuk menentukan klasifikasi karakteristik calon mahasiswa.

3.3. *Algoritma Naïve Bayes*

Klasifikasi membentuk model data yang mampu untuk membedakan data dalam kelas yang berbeda menggunakan aturan dan fungsi tertentu. Salah satu algoritma yang digunakan pada teknik klasifikasi adalah Algoritma *Naïve Bayes*. Pada awalnya, teori tersebut diadopsi dari teori yang dikemukakan oleh Thomas Bayes yaitu perhitungan peluang yang akan terjadi berdasarkan pengalaman yang sudah terjadi [4].

Teori Bayes dituliskan dalam perhitungan seperti berikut:

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

Berdasarkan rumus tersebut dapat dijelaskan sebagai berikut:

1. P(A|B) merupakan peluang terjadinya kejadian A dengan syarat kejadian B telah terjadi
2. P(A|B) merupakan peluang terjadinya kejadian B dengan syarat kejadian A telah terjadi
3. P(A) merupakan peluang terjadinya kejadian A, tanpa pengaruh kejadian lain
4. P(B) merupakan peluang terjadinya kejadian B, tanpa pengaruh kejadian lain

Klasifikasi pada Bayes menjelaskan bahwa variabel B merupakan sekumpulan variabel pada data, sedangkan variabel A merupakan kelas pada data. Apabila kelas pada suatu data memiliki hubungan non-deterministik dengan variabel-variabel maka A dan B dapat dianggap variabel acak. Naïve dalam hal ini menyempurnakan rumus Bayes dengan asumsi variabel yang mempengaruhi bersifat mandiri [4]. Teori tersebut dimanfaatkan oleh Naïve untuk memperhitungkan sebuah peluang dengan variabel bebas. Adapun alur dari metode *Naïve Bayes* adalah sebagai berikut [4]:

1. Menghitung nilai peluang dari setiap kelas.
2. Menghitung nilai peluang P(Bi | A) atau peluang bersyarat tiap-tiap variable terhadap kelas.
3. Menentukan label ke arah positif atau negatif dari peluang yang dihitung. Teori Naïve Bayes dinyatakan dalam perhitungan sebagai berikut:

$$P(A|B) = \frac{P(A)\pi_{i=1}^{d} P(B_i|A)}{P(B)}$$

Berdasarkan rumus tersebut terdapat perbedaan sedikit dengan teori Bayes yaitu pada $P(B_i|A)$ yang berarti peluang terjadinya setiap variabel B_i dengan syarat A. Peluang B bersifat konstan untuk setiap kejadian A, sehingga nilai peluang B dapat dihilangkan dari persamaan. Sedangkan untuk melakukan penentuan kelas dihitung dari nilai maksimal $P(A_k \mid B)$ untuk setiap kelas yang terdapat pada data [4]. Perhitungan tersebut dinyatakan dalam persamaan sebagai berikut:

$$C = \underset{k \in \{1,\dots,K\}}{argmax} P(A_k)\pi_{i=1}^{d} P(B_i|A)$$

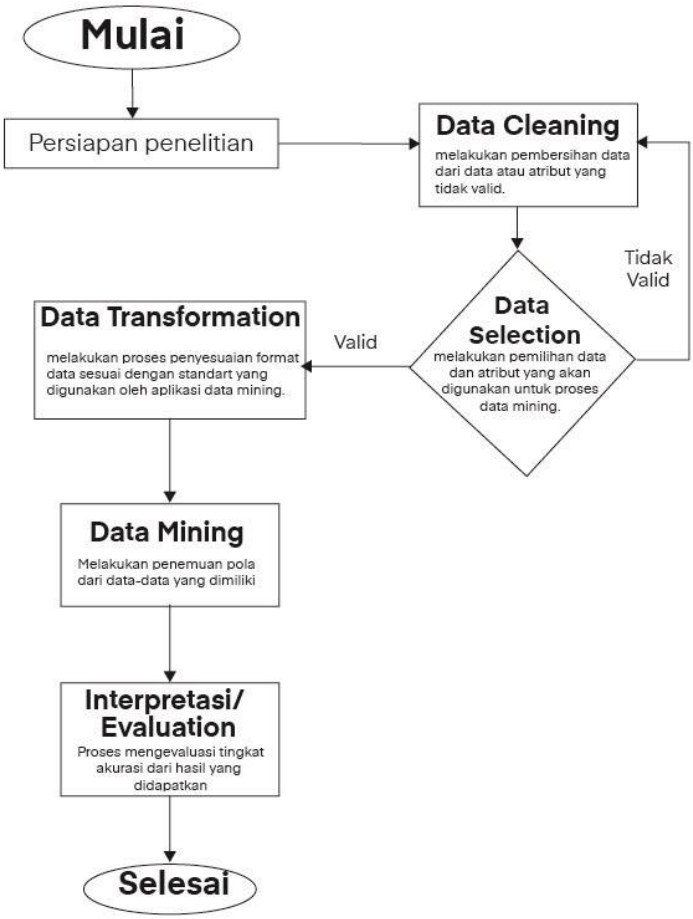
K dalam persamaan tersebut merupakan himpunan kelas yang terdapat pada data yang ada. Metode ini cenderung lebih mudah dipahami karena parameter yang digunakan dapat dipakai secara berulang, bahkan Ketika *dataset* bertambah besar.

Naïve Bayes merupakan algoritma yang mudah diimplementasikan pada banyak kasus dengan hasil yang baik, namun kelemahan algoritma tersebut tidak dapat menggambarkan sebuah keterkaitan antar variabel. Algoritma *Naïve Bayes* mampu menganalisis kelompok dengan kemungkinan tertinggi dari kelompok-kelompok yang telah dibuat dan kondisi yang diberikan. Menurut Yuda Septian Nugroho, Naïve Bayes merupakan algoritma yang mampu memberikan hasil dengan bukti hasil akurasi dan kecepatan yang tinggi ketika diaplikasikan pada database [4].

Berdasarkan penelitian dan penjelasan dari beberapa sumber, Naïve *Bayes* dapat menjadi salah satu algoritma yang cocok dengan penelitian ini dikarenakan tujuan dari algoritma tersebut sama dengan tujuan penelitian ini yaitu untuk menentukan sebuah probabilitas dan pengklasifikasian data yang dalam hal ini diimplementasikan pada data calon mahasiswa dan mahasiswa yang telah berhasil mendaftar.

4. METODE PENELITIAN

Desain penelitian yang digunakan menggunakan model KDD. Gambar 1 merupakan gambaran alur penelitian ini dilakukan dari awal hingga akhir.



Gambar 1
Metode Penelitian

Tahapan pada KDD yang diimplementasikan pada penelitian ini adalah sebagai berikut:

1. *Data Cleanning*
   Data cleaning merupakan proses membuang duplikasi data, memeriksa data yang tidak konsisten, dan memperbaiki kesalahan pada data, seperti kesalahan penulisan.
2. *Data Selection*
   Pada tahap ini dilakukan penyesuaian kebutuhan data dengan tujuan bisnis perusahaan. Pemahaman akan hasil akhir dibutuhkan untuk mengerti kebutuhan data dan atribut yang akan diolah oleh Data Mining. Data yang digunakan adalah data yang didapat dari formulir Pendaftaran Mahasiswa Baru Jalur Undangan dan juga data hasil proses pengajuan beasiswa. Data pendukung lainnya didapatkan dari hasil Survei pada mahasiswa Institusi Pendidikan X, data yang dibutuhkan tersebut merupakan data eknomi keluarga, data demografi seperti Hobi dan bidang minat, data alasan memilih Institusi Pendidikan X sebagai kampus dan juga data keberadaan anggota keluarga atau saudara di Institusi Pendidikan X sebelumnya.
3. *Data Transformation*
   Pada penelitian ini dilakukan proses penyesuaian format data yang akan digunakan pada perangkat lunak Data Mining. Penelitian ini menggunakan aplikasi/ *software*

yang mendukung Data Mining yaitu DBFViewer dan Microsoft Excell dalam melakukan transformasi data.

4. *Data Mining*
   Pada tahap ini mencoba menemuka pola-pola dari data yang ada dan menentukan karakteristik dari data yang telah diolah. Algoritma yang digunakan untuk menemukan pola dan membentuk klasifikasi adalah Naïve Bayes. Klasifikasi ditentukan untuk menemukan kesamaan data dari data set mahasiswa Ketika pertama kali mendaftar dan Ketika sudah menjadi mahasiswa di Institusi Pendidikan X. Software yang digunakan adalah WEKA, pada alat bantu tersebut menawarkan banyak teknik Data Mining seperti: preprocessing data, klasifikasi, regresi, pengelompokan, aturan asosiasi, dan visualisasi, dengan antarmuka yang mudah. Pada penelitian ini akan menggunakan WEKA Toolkit 3.8.4
5. *Interpretasion/ Evaluation*
   Setelah menemukan pola dan karakteristik yang ada, maka dapat dievaluasi akurasi dari data dan menentukan probabilitas untuk setiap kategori untuk mengetahui kemungkinan calon mahasiswa untuk menjadi mahasiswa.

5. HASIL PENELITIAN

5.1. *Data Cleanning*

Langkah pertama adalah mencoba memasukan data ke dalam pemrosesan data di dalam aplikasi WEKA, beberapa atribut memiliki pola yang terlalu beragam/ Data Unik dan data bersifat Null sehingga tidak ditemukan pola dengan metode klasifikasi Naïve Bayes. Atribut data yang dimaksud seperti: Nomor USM, Nama, Kewarganegaraan, golongan darah, Nomor telepon, Nomor HP, Alamat 1, Alamat 2, Alamat Bandung 1, Alamat Bandung 2, Kodepos 1, Kodepos 2, Kodepos asal SMU, Sumbangan tambahan, Pilihan cadangan dan tanggal Lahir. Setelah melihat lebih jelas dan detail bahwa atribut sudah sesuai dengan kebutuhan informasi, maka pembersihan dilakukan pada baris data yang hilang agar kualitas data lebih baik. Penjabaran atribut-atribut tersebut dirangkum dalam Tabel 2.

Tabel 2
Verifikasi Kualitas Data

Atribut	Missing Value	Distinct	Unique
Gelombang	0	9	0
Pilihan Utama	37	28	2
Asal Kota	87	354	191
Asal SMU	35	1172	918
USM	0	2	0
Beasiswa yang diajukan	0	4	0
Hasil Beasiswa	0	7	0
Daftar Ulang	0	2	0

Berdasarkan data tersebut maka akan dilakukan pembersihan Kembali terhadap data bernilai null atau kosong pada data calon mahasiswa Institusi Pendidikan X sebagai berikut:

1. Data kosong pada atribut pilihan utama program studi sebanyak 37 data akan dihapus, maka jumlah data akan menjadi 2829 data.

2. Data kosong pada atribut asal kota sebanyak 87 data akan dihapus maka jumlah data akan menjadi 2742
3. Data kosong pada atribut jenis kelamin sebanyak 67 data akan dihapus maka jumlah data akan menjadi 2675 data.

Data kosong pada atribut asal SMU sebanyak 35 data akan dihapus maka jumlah data akan menjadi 2640 data

5.2. *Data Selection*

Setelah melakukan pembersihan, pada tahap ini data akan disiapkan untuk dilakukan pengolahan menggunakan *Data Mining*. Data Calon Mahasiswa yang didapatkan dari Institusi Pendidikan X memiliki 28 atribut setelah digabungkan dan dilakukan normalisasi dan terpilih menjadi sebanyak 8 atribut yang akan digunakan untuk melihat pola yang ada.

5.3. *Data Transformasi*

Pada tahap ini data tersebut disesuaikan dengan format dari sistem WEKA sebagai alat untuk melakukan *Data Mining*. Data yang didapatkan berformat dbf sedangkan pengolahan dilakukan dalam format csv, maka perlu proses melakukan transformasi format data. Transformasi format data dilakukan menggunakan aplikasi bernama DBFViewer, pada aplikasi DBFViewer terdapat fitur untuk mengekstrak data ke dalam bentuk Excel. Selain itu pada tahap ini dilakukan penyesuaian data dengan menggunakan Microsoft Excel seperti melakukan sort dan filter.

Pada tahap ini juga dilakukan normalisasi data dari beberapa tabel, yaitu tabel data formulir dan data tabel Daftar Ulang, sedangkan data-data pendukung dianalisis terpisah. Pengkodean pada data yang memiliki format yang tidak sesuai seperti penggunaan angka untuk kode diubah menjadi huruf.

5.4. *Data Mining*

Setelah data CSV dimasukan ke dalam sistem WEKA, dilakukan klasifikasi dengan memilih *classifiy* dari *package Bayes* yaitu *Naïve Bayes*. *Test Option* menggunakan *Percentage Split* dikarenakan data pertahun yang banyak salah satunya adalah data tahun 2020 yaitu sekitar 26% dari keseluruhan data, maka akan dilakukan *Percentage Split* dengan 74% data latih dan 26% data uji.

Implementasi tahapan tersebut pada Weka dan *memberikan* hasil pada Gambar 2.

```
=== Summary ===

Correctly Classified Instances         553               80.6122 %
Incorrectly Classified Instances       133               19.3878 %
Kappa statistic                          0.5328
Mean absolute error                      0.1997
Root mean squared error                  0.3667
Relative absolute error                 52.4543 %
Root relative squared error             85.0356 %
Total Number of Instances              686

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.781    0.186    0.579      0.781    0.665      0.545    0.897     0.683     Y
               0.814    0.219    0.919      0.814    0.864      0.545    0.897     0.966     T
Weighted Avg.  0.806    0.211    0.835      0.806    0.815      0.545    0.897     0.896

=== Confusion Matrix ===

   a   b   <-- classified as
 132  37 |   a = Y
  96 421 |   b = T
```

Gambar 2
Classifier Summary

Pada Gambar 2 dapat dilihat bahwa nilai *error* yang diperoleh adalah 19.3878% yaitu sekitar 133 data dari 786 data uji dan data berhasilnya adalah 80.6122% yaitu sekitar 553 data dari 786 data uji. Berdasarkan perhitungan melalui fitur Classifiy dengan mengikuti urutan pada Information Gain maka atribut yang memberikan pola paling akurat untuk menentukan seorang calon mahasiswa melakukan daftar ulang adalah :

1. *Gelombang* untuk melihat waktu pendaftaran yang tepat
2. Agama untuk melihat karakteritik lingkungan dan social calon mahasiswa
3. USM untuk melihat efektifitas pemberian beasiswa dengan syarat ujian
4. Asal Kota untuk melihat karakteristik dari segi ekonomi, social, dll
5. Asal SMU untuk melihat karakteritik dari segi ekonomi

Setelah dicoba kembali prosesnya maka dipilih beberapa model pola data yang memiliki data cukup tinggi seperti pada Tabel 3.

Tabel 3
Karakteristik Calon Mahasiswa 2016-2020

Pola (Gelombang- Asal Kota-Agama- Asal SMU-USM)	Margin Prediction	Jumlah Data
Z - BANDUNG - B - SMAK GAMALIEL - T	0.334553	93
Z - BANDUNG - B - SMAK KALAM KUDUS - T	0.374817	34
Z - BANDUNG - B - SMAK 3 BINA BAKTI - T	0.309332	23
Z - BANDUNG - C - SMA SANTA MARIA 1 - T	0.05598	21
Z - BANDUNG - B - SMAK 2 BPK PENABUR - T	0.490428	20
Z - CIMAHI - B - SMA SANTA MARIA 3 - T	0.560147	11
Z - BANDUNG - C - SMAK GAMALIEL - T	0.148736	9

5.5. *Data Evaluation*/ Interpretasi

Berdasarkan *confusion matrix* data tersebut didapatkan hasil dari sistem yaitu:

1. *True Positive Rate* (hasil prediksi yang nilainya sama-sama benar dengan sebenarnya) sebanyak 136 data
2. *False Negative Rate* (hasil prediksi salah, tetapi hasil yang sebenarnya bernilai Benar) sebanyak 33 data
3. *True Negative Rate* (hasil prediksi salah, tetapi hasil yang sebenarnya bernilai salah) sebanyak 436 data
4. *False Positive Rate* (hasil yang benar, tetapi hasil yang sebenarnya salah) sebanyak 81 data

6. KESIMPULAN

Kesimpulan dari penelitian ini adalah sebagai berikut:

a. Beasiswa Penuh yang diberikan dapat diberikan lebih banyak untuk karakteristik calon mahasiswa yang memiliki prestasi lebih di bidang akademik tanpa harus mengikuti ujian kembali, hal itu dapat dilihat dari nilai yang paling memiliki tingkat akurasi klasifikasi class USM adalah T yaitu sekitar 83,3%
b. Program-program Beasiswa lebih baik diberikan pada gelombang awal (Z) atau sekitar dibulan Agustus-Oktober.

c. Karakteristik gabungan yang paling tinggi dengan presentase 33,3% adalah calon mahasiswa yang mendaftar di gelombang awal (Z) dengan tingkat ekonomi menengah yaitu yang bersekolah di sekolah SMAK Gamaliel Bandung dan memiliki prestasi sehingga tidak perlu mengikuti ujian untuk mendapatkan beasiswanya.

DAFTAR PUSTAKA

[1] Buttle, Francis and Stan Maklan, 2015, *Customer Relationship Management: Concepts and technologies,* London: Butterworth-Heinemann, an imprint of Elsevier.

[2] Fayyad, Usama, Gregory Piatetsky-Shapiro, dan Padhraic Smyth, 1996, *From Data Mining to Knowledge Discovery in Databases,* USA: American Association for Artificial Intelligence.

[3] Hananto, Valentinus Robu, dkk, 2017, *Perancangan Analytical CRM untuk Mendukung Segmentasi Pelanggan di Institusi Pendidikan,* Surabaya: Institut Bisnis dan Informatika Stikom.

[4] Nugroho, Yuda Septian, 2013, *Data Mining Menggunakan Algoritma Naïve Bayes Untuk Klasifikasi Kelulusan Mahasiswa Universitas Dian Nuswantoro,* Indonesia: Universitas Dian Nuswantoro

[5] Nada, Riri Devita, 2018, *Perbandingan Kinerja Metode Naive Bayes dan K-Nearest Neighbor untuk Klasifikasi Artikel Berbahasa Indonesia,* Indonesia: Universitas Negeri Malang.

[6] Oluseye, Ogunnaike Olaleke, dkk, 2014, *Customer Relationship Management Approach and Student Satisfaction in Higher Education Marketing,* Nigeria.

[7] Oliveira, Vera Lucia Migueis, 2012, *Analytical Customer Relationship Management in Retailing Supported by Data Mining Techniques,* Prancis: Universidade do Porto.

[8] Salim, Ahmad, dkk, 2017, *Predicting Student Enrollment Based on Student and College Characteristics,* Mexico: University of New Mexico

[9] Turban, E., McLean, E., and Wetherbe, J.,2006, *Information Technology for Management: Transforming Organisations in the Digital Economy,* New York: John Wiley