Implementasi Teorema *Naïve Bayes* Pada Prediksi Prestasi Mahasiswa

Rofilde Hasudungan*[1], Wawan Joko Pranoto[2]
[1,2]Fakultas Sains dan Teknologi; Jalan Juanda; Telp. 0541-748511 Fax. 0541-766832;
Universitas Muhammadiyah Kalimantan Timur
e-mail: ***[1]rofilde@umkt.ac.id**, [2]wjp337@umkt.ac.id

Abstrak

Informasi tentang prestasi akademik mahasiswa merupakan hal yang sangat penting untuk diketahui. Prediksi mengenai prestasi akademik mahasiswa secara dini dapat menentukan tindakan-tindakan yang diperlukan untuk meningkatkan prestasi mahasiswa yang diprediksi memiliki prestasi yang rendah, sehingga dikemudian hari dapat memiliki prestasi yang baik. Namun, informasi untuk mengetahui prestasi akademik mahasiswa sangat sulit dilakukan. Penelitian ini mengungkap faktor keluarga terhadap prestasi menggunakan algoritma Naïve Bayes dan enam belas parameter yang terlibat. Data yang digunakan dalam penelitian ini adalah 40 data mahasiswa. Hasil penelitian menunjukkan bahwa algoritma Naïve Bayes dapat memprediksi prestasi akademik mahasiswa dengan tingkat akurasi sebesar 77,5%.

Kata kunci— *Cross Validation, Mahasiswa, Naïve Bayes, Prestasi, Prediksi*

1. PENDAHULUAN

Kemampuan untuk memprediksi prestasi mahasiswa merupakan suatu hal yang sangat penting. Informasi tentang prestasi akademik mahasiswa sejak dini dapat mencegah mahasiswa yang diprediksi memiliki prestasi yang rendah untuk gagal dalam perkuliahan. Informasi ini juga sangat penting bagi manajemen Universitas untuk memonitoring proses pembelajaran dan melakukan kebijakan yang diperlukan untuk meningkatkan prestasi mahasiswa. Bagi dosen, informasi ini sangat berguna untuk mengevaluasi teknik, kualitas dan pendekatan dalam pembelajaran di dalam perkuliahan. Namun, memprediksi prestasi mahasiswa merupakan masalah yang rumit, hal ini disebabkan terdapat banyak faktor yang terlibat, seperti psikologi, geografi, sosial-ekonomi, gaya mengajar dosen dan lingkungan akademik.

Dari sisi penjaminan mutu pendidikan, umumnya semua Universitas di Indonesia menerapkan suatu mekanisme untuk mengevaluasi dan memonitoring proses belajar mengajar. Dari sisi pengajaran pada umumnya evaluasi dilakukan terhadap hasil pembelajaran yang dilakukan oleh dosen terhadap mata kuliah yang diampu. Evaluasi yang dilakukan terhadap dosen dilakukan dengan menggunakan kuesioner yang akan diisi oleh mahasiswa. Pengisian kuesioner umumnya dilakukan di akhir semester dan hasilnya dalam bentuk skor rata-rata penilaian mahasiswa terhadap dosen pada suatu mata kuliah. Pendekatan seperti ini memiliki banyak kekurangan, diantaranya adalah tidak menjawab faktor-faktor apa saja yang berpengaruh terhadap prestasi mahasiswa dan pengambilan data yang dilakukan di akhir semester tidak dapat membantu mahasiswa maupun dosen dalam mengevaluasi teknik dan pendekatan pembelajaran pada semester tersebut yang dapat mencegah mahasiswa gagal dalam perkuliahan. Oleh karena itu, diperlukan pendekatan dan metode baru untuk menanggulangi permasalahan tersebut.

Data mining merupakan suatu alat yang dapat digunakan untuk mengungkapkan informasi didalam data yang besar. Data mining yang juga disebut sebagai *Knowledge in*

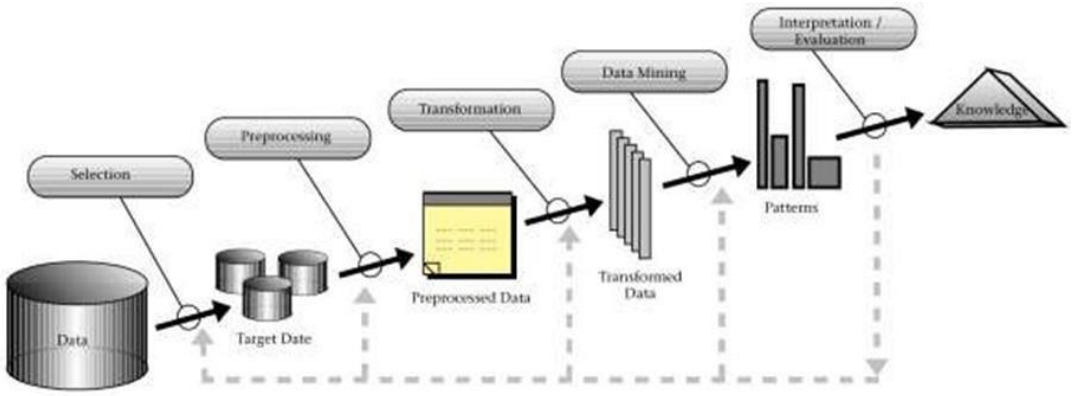
Discovery Database (KDD) telah digunakan diberbagai bidang seperti bisnis, *e-commerce*, astronomi, geografi, kesehatan dan pendidikan. Di dalam data mining terdapat beberapa teknik yang digunakan untuk mengungkap informasi dalam suatu data seperti teknik klasifikasi (*classification*), pengelompokan (*clustering*) dan assosiasi (*Association*). Penerapan teknik data mining telah dilakukan oleh beberapa penelitian diantaranya pengelompokkan segmentasi pelanggan [1], rekomendasi tugas akhir [2], klasifikasi masyarakat miskin, dosen berprestasi, beasiswa PPA [3]–[5] dan lain sebagainya [6]–[8].

Pada penelitian ini akan membahas penggunaan data mining khususnya teknik klasifikasi dengan menggunakan *Naïve Bayes* untuk memprediksi prestasi mahasiswa. Teorema *Naïve Bayes* sendiri telah digunakan untuk menganalisis opini masyarakat tentang penanganan penyakit difteri oleh pemerintah dengan persentase sebesar 94,5% opini masyarakat adalah negatif [9]. Lebih lanjut, penerapan Teorema *Naïve Bayes* dapat digunakan untuk mengidentifikasi hama dan penyakit pada anggrek hitam [10], [11]. Penelitian ini diharapkan dapat mengungkapkan penggunaan Teorema *Naïve Bayes* dalam memprediksi data pendidikan.

2. METODE PENELITIAN

2.1 Data Mining Pada Pendidikan

Peningkatan penggunaan teknologi informasi menghasilkan data yang sangat besar. Data-data ini menumpuk, tetapi tidak memiliki nilai, padahal data-data yang menumpuk dapat mengandung informasi yang sangat berharga. Data mining (KDD) merupakan suatu alat yang digunakan untuk mengungkapkan informasi yang tersembunyi dalam tumpukan data. Di dalam melakukan penambangan terhadap data, terdapat beberapa teknik data mining yaitu klasifikasi, pengelompokan, anomali, *detection* dan assosiasi. Gambar 1 menunjukkan proses yang umum digunakan untuk mengungkap informasi atau pengetahuan (*knowledge*) dari suatu penambangan data (*data mining*).



Gambar 1 Proses pengungkapan data pada *data mining*

Data mining telah banyak dimanfaatkan di berbagai bidang, salah satunya dibidang pendidikan. Di dalam dunia pendidikan istilah yang lebih umum dinamakan *Data Mining in Education* (EDM) yang mengkombinasikan data mining dalam mengungkap informasi-informasi penting dalam data-data yang berkaitan dengan pendidikan. Data mining dalam dunia pendidikan banyak digunakan untuk mengelompokan mahasiswa berdasarkan suatu karakteristik tertentu misalkan gaya belajar atau secara kedekatan personal, melakukan prediksi terhadap prestasi mahasiswa dan melakukan pengungkapan faktor-faktor yang mempengaruhi prestasi mahasiswa.

Penelitian terkait penggunaan data mining dalam dunia pendidikan sudah banyak dilakukan diantaranya menggunakan *K-Means* untuk mengetahui pola keterikatan antara faktor akademik untuk memprediksi prestasi mahasiswa berdasarkan catatan prestasi akademik [12]. Selanjutnya, penelitian yang menerapkan algoritma pengelompokan *Fuzzy C-Means* untuk mengelompokkan mahasiswa berdasarkan Indeks Prestasi Kumulatif (IPK) dan lama kelulusan. Pengelompokkan ini bertujuan untuk membagi mahasiswa kedalam suatu interval data yang dapat diklasifikasikan ke dalam 4 kelompok utama yaitu jelek, tidak bagus, sangat bagus dan bagus [13]. Penelitian lainnya mengungkapkan bahwa tingkat kehadiran dan Indeks Prestasi (IP) mahasiswa pada semester merupakan faktor penting yang berpengaruh terhadap prestasi mahasiswa menggunakan dua teknik data mining yaitu *Naïve Bayes* dan *Decision Tree.* Penelitian ini juga mengungkapkan bahwa algoritma *Naïve Bayes* memiliki tingkat akurasi yang tinggi dibandingkan *Decision Tree* dalam memprediksi prestasi mahasiswa [14]. Teknik data mining lainnya seperti algoritma C4.5, *Support Vector Machine* (SVM) dan *Artificial Neural Network* telah banyak digunakan dalam dunia pendidikan untuk mengungkapkan pola maupun informasi yang diperlukan dalam dunia pendidikan [15], [16].

2.2 Teorema Naïve Bayes

Teorema *Naïve Bayes* adalah salah satu metode untuk mengatasi ketidakpastian. Metode ini dapat dikatakan memprediksi peluang di masa depan berdasarkan pengalaman dimasa sebelumnya [10]. Klasifikasi *Naïve Bayes* diasumsikan bahwa ada atau tidak ciri tertentu dari sebuah kelas tidak ada hubungannya dengan kelas lainnya. Jumlah data pelatihan (*training*) yang dibutuhkan oleh metode *Naïve Bayes* sedikit, dimana hal ini merupakan keunggulan dari metode ini. Data *training* digunakan untuk menentukan etimasi parameter yang diperlukan dalam proses pengklasifikasian, karena yang diasumsikan sebagai variabel *independent*, maka hanya varians dari suatu variabel dalam sebuah kelas yang dibutuhkan untuk menentukan klasifikasi, bukan keseluruhan dari matriks kovarians. Persamaan (1) merupakan persamaan dari metode Naïve Bayes [17].

$$P\ (H|E) = \frac{P(E|H) \times P(H)}{P(E)} \qquad (1)$$

Keterangan:
P(H|E) : Probabilitas akhir bersyarat (*conditional probability*) suatu hipotesis H terjadi jika diberikan bukti (*evidence*) E terjadi.
P(E|H) : Probabilitas sebuah bukti E akan memengaruhi hipotesis H.
P(H) : Probabilitas awal hipotesis H terjadi tanpa memandang bukti apapun.
P(E) : Probabilitas awal bukti E terjadi tanpa memandang hipotesis/bukti yang lain.

2.3 Cross Validation

K-fold adalah salah satu metode *Cross Validation* yang populer dengan melipat data sebanyak *k* dan mengulangi (*iterasi*) eksperimennya sebanyak *k* juga. Pada pengujian didalam penelitian ini menggunakan *k = 10*. Lebih lanjut, ditentukan mana yang termasuk data *training* dan mana yang termasuk data *testing* dengan perbandingan 9:1. Pengujian menggunakan data yang sudah dipartisi akan diulang sebanyak 10 kali (*k* =10) dengan posisi data *testing* berbeda disetiap iterasinya. Misalkan iterasi pertama data tes pada posisi awal, iterasi kedua data testing di posisi kedua begitu seterusnya [18].

2.4 Atribut dan Data Penelitian

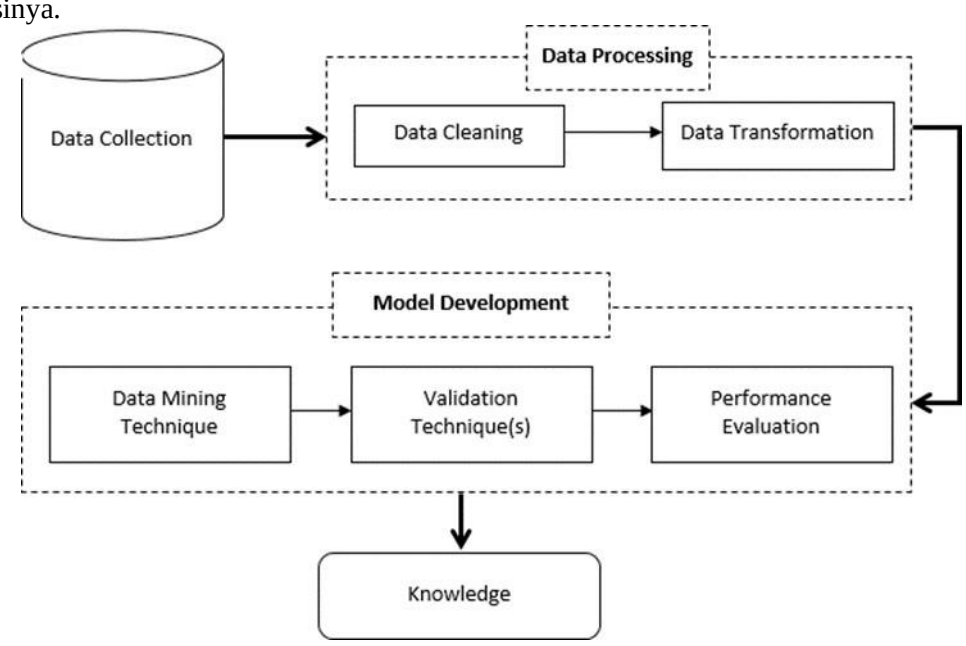
Data yang digunakan dalam penelitian ini mengambil data dari mahasiswa aktif semester ke 3, angkatan 2018 dan data-data yang terkait dengan keluarga mahasiswa. Adapun indikator yang digunakan pada penelitian ini ditunjukan pada Tabel 1.

Tabel 1 Indikator prediksi

Parameter	Keterangan
Umur	Umur Mahasiswa
Jenis Kelamin	Jenis kelamin mahasiswa
Kategori Tempat Tinggal	Status Tempat Tinggal: 1) Bersama Keluarga, 2) Orang Tua, 3) Kos, 4) Kontrak Bersama Teman
Jarak	Jarak dari tempat tinggal ke Kampus
Jenis Pendidikan SLTA	Jenis Pendidikan terakhir SMA/SMA/MA
Status Pendidikan Sebelumnya	Status institusi pendidikan sebelumnya: (1) Negri (2) Swasta
PekerjaanAyah	Pekerjaan Ayah
Pekerjaan Ibu	Pekerjaan Ibu
Pendidikan Ayah	Level pendidikan ayah
Pendidikan Ibu	Level pendidikan ibu
Status Orang Tua	Status pernihakan orang tua,
Pengahasilan Ayah	Penghasilan ayah
Pengahasilan Ibu	Penghasil ibu
Jumlah Anggota Keluarga	Jumlah anggota keluarga dalam satu ruma
JumlahKakak	Jumlah kakak
Jumlah Adik	Jumlah Adik

3.2 Model Klasifikasi

Pada penelitian ini mengusulkan model klasifikasi seperti ditunjukkan oleh Gambar 2. Pada model ini, data yang telah didapatkan akan dilakukan proses *data cleaning* yaitu dengan melakukan atau pembersihan data anomali. Data selanjutnya dianalisa dengan menggunakan algoritma *Naïve Bayes*, kemudian di validasi menggunakan c*ross validation* untuk mendapatkan akurasinya.



Gambar 2 Desain model klasifikasi dengan *Naïve Bayes*

3. HASIL DAN PEMBAHASAN

Penelitian ini menggunakan *tools RapidMiner* untuk menjalankan model penelitian seperti pada Gambar 2. *Outlier detection* yang digunakan adalah pendekatan yang dilakukan oleh Ramaswamy dkk, dimana pendekatan ini mengukur berdasarkan jarak (*eucledien distance*). Pada penelitian ini terdapat dua buah parameter yang diatur yaitu *number of neighbors* dan *number of outliers* dengan nilai masing-masing 7 dan 5. Nilai ini diperoleh dengan beberapa kali percobaan dengan hasil akurasi yang terbaik. Data yang telah bebas dari data anomali, kemudian di analisa menggunakan *Naïve Bayes*. Lebih lanjut, hasil yang didapatkan kemudian di validasi menggunakan *cross-validation* dengan parameter *number of folds* sebesar 10. Hasil yang didapatkan terlihat pada Tabel 2.

Tabel 2 Performa Prediksi Model

	True Baik	True Sangat Baik	True Cukup	Class Precision
Pred. Baik	26	3	3	80.25%
Pred. Sangat Baik	1	1	0	50.00%
Pred. Cukup	0	1	1	0.00%
Class Recall	96.30%	20.00%	0.00%	

Berdasarkan Tabel 2, akurasi dari model yang diusulkan adalah 77.5%. Akurasi ini meningkat dari percobaan sebelumnya yang tidak menggunakan deteksi anomali, dimana pada percobaan sebelumnya tingkat akurasinya hanya 69% seperti ditunjukan pada Tabel 3.

Tabel 3 Performa Prediksi Model tanpa deteksi anomali

	True Baik	True Sangat Baik	True Cukup	Class Precision
Pred. Baik	26	4	4	76.47%
Pred. Sangat Baik	3	1	0	25.00%
Pred. Cukup	1	1	0	0.00%
Class Recall	86.67%	16.67%	0.00%	

4. KESIMPULAN

Analisis data mahasiswa merupakan hal yang sangat penting untuk mengetahui pengaruh terhadap prestasi akademik. Penelitian ini melakukan analisa terhadap data latar belakang keluarga mahasiswa terhadap prestasi mahasiswa yang diukur dari nilai IPK. Data yang digunakan dalam penelitian ini sebanyak 40 data mahasiswa, kemudian setelah menggunakan teknik *outlier detection*, terdapat 5 data yang dianggap anomali. Model yang diusulkan pada penelitian ini berbasis *Naïve Bayes* sebagai *classifier,* sehingga setiap parameter dianggap sama pentingnya. Dari penelitian yang telah dilakukan, hasil analisa menunjukan bahwa model yang diusulkan memiliki tingkat akurasi sebesar 77,5%, dan hasil yang lebih rendah sebesar 69% bila tidak menggunakan *outlier detection*.

5. SARAN

Saran untuk penelitian selanjutnya adalah meningkatkan akurasi dari model ini dengan menerapkan beberapa pendekatan seperti pemilihan atribut yang tepat, atau menambahkan data yang lebih besar. Selain memilih indikator yang tepat untuk meningkatkan akurasi, pengungkapan indikator dan relasi antar indikator menjadi topik yang sangat penting. Dengan mengetahui indikator yang mempengaruhi prestasi mahasiswa, selanjutnya dapat dilakukan tindakan yang lebih konkret.

DAFTAR PUSTAKA

[1] N. Puspitasari, J. A. Widians, and N. B. Setiawan, "Segmentasi pelanggan menggunakan algoritme bisecting k-means berdasarkan model recency, frequency, dan monetary (RFM)," *J. Teknol. dan Sist. Komput.*, vol. 8, no. 2, 2020.

[2] Haviluddin, S. J. Patandianan, G. M. Putra, and H. S. Pakpahan, "Implementasi Metode K-Means untuk Pengelompokkan Rekomendasi Tugas Akhir," *Inform. Mulawarna J. Ilm. Ilmu Komput.*, vol. 16, no. 1, 2021.

[3] H. Annur, "Klasifikasi Masyarakat Miskin Menggunakan Metode Naive Bayes," *Ilk. J. Ilm.*, vol. 10, no. 2, pp. 160–165, 2018.

[4] I. Purnamasari and K. Afnisari, "Performansi Klasifikasi Dosen Berprestasi Menggunakan Metode Naive Bayes Classifier," *Paradig. Komput. dan Inform.*, vol. 20, no. 2, pp. 45–50, 2018.

[5] S. Adi, "Implementasi Algoritma Naive Bayes Classifier Untuk Klasifikasi Penerima Beasiswa PPA Di Universitas Amikom Yogyakarta," *J. Mantik Penusa*, vol. 22, no. 1, 2018.

[6] A. A. Rahman and Y. I. Kurniawan, "Aplikasi Klasifikasi Penerima Kartu Indonesia Sehat Menggunakan Algoritma Naïve Bayes Classifier," *J. Teknol. dan Manaj. Inform.*, vol. 4, no. 1, 2018.

[7] H. Mustofa and A. A. Mahfudh, "Klasifikasi Berita Hoax Dengan Menggunakan Metode Naive Bayes," *Walisongo J. Inf. Technol.*, vol. 1, no. 1, pp. 1–12, 2019.

[8] I. Ramadhan and K. Kurniawati, "Data Mining untuk Klasifikasi Penderita Kanker Payudara Berdasarkan Data dari University Medical Center Menggunakan Algoritma Naïve Bayes," *JURIKOM (Jurnal Ris. Komputer)*, vol. 7, no. 1, pp. 21–27, 2020.

[9] A. Sholihin, H. Haviluddin, N. Puspitasari, M. Wati, and I. Islamiyah, "Analisis Penyakit Difteri Berbasis Twitter Menggunakan Algoritma Naïve Bayes," *Sains, Apl. Komputasi dan Teknol. Inf.*, vol. 1, no. 1, 2019, doi: 10.30872/jsakti.v1i1.2215.

[10] J. A. Widians, N. Puspitasari, and A. A. M. Putri, "Penerapan Teorema Bayes dalam Sistem Pakar Anggrek Hitam," *Inform. Mulawarman J. Ilm. Ilmu Komput.*, vol. 15, no. 2, 2020, doi: 10.30872/jim.v15i2.4604.

[11] J. A. Widians, N. Puspitasari, H. S. Pakpahan, E. Budiman, and F. Alameka, "Identification Pests and Diseases of the Borneo Black Sweet in Tropical Rainforest," *J. Phys. Conf. Ser.*, 2021, doi: 10.1088/1742-6596/1844/1/012007.

[12] G. S. Gowri, R. Thulasiram, and M. A. Baburao, "Educational data mining application for estimating students performance in weka environment," in *IOP Conference Series: Materials Science and Engineering*, 2017, vol. 263, no. 3, p. 32002.

[13] R. Rosadi, R. Sudrajat, B. Kharismawan, and Y. A. Hambali, "Student academic performance analysis using fuzzy C-means clustering," in *IOP Conference Series: Materials Science and Engineering*, 2017, vol. 166, no. 1, p. 12036.

[14] A. U. Khasanah, "A comparative study to predict student's performance using educational data mining techniques," in *IOP Conference Series: Materials Science and Engineering*, 2017, vol. 215, no. 1, p. 12036.

[15] M. A. Al-Barrak and M. Al-Razgan, "Predicting students final GPA using decision trees: a case study," *Int. J. Inf. Educ. Technol.*, vol. 6, no. 7, p. 528, 2016.

[16] Y. C. Giap, N. Leonardi, B. Waseso, and R. Rahim, "Data Mining of Family, School, and Society Environments Influences to Student Performance," in *IOP Conference Series: Materials Science and Engineering*, 2018, vol. 420, no. 1, p. 12090.

[17] N. R. Indraswari and Y. I. Kurniawan, "Aplikasi prediksi usia kelahiran dengan metode naive bayes," *Simetris J. Tek. Mesin, Elektro dan Ilmu Komput.*, vol. 9, no. 1, pp. 129–

138, 2018.

[18] O. Dwiraswati and K. N. Siregar, “Analisis Sentimen Pada Twitter Terhadap Penggunaan Antibiotik di Indonesia dengan Naive Bayes Classifier,” *Media Inf.*, vol. 15, no. 1, pp. 1–9, 2019.