

Penerapan Algoritma Naive Bayes Dan *K-Means Clustering* Untuk Prediksi Dan Segmentasi Mahasiswa Berdasarkan Kinerja Akademik Dan Lama Kelulusan

Andri Rachmatullah¹, Kusrini²

^{1,2}Universitas Amikom Yogyakarta

Corresponding Author: andrirachmatullah90@gmail.com, kusrini@amikomac.id

Published: Agust , 2025

ABSTRAK

Penelitian ini bertujuan untuk mengelompokkan mahasiswa berdasarkan data akademik menggunakan algoritma K-Means Clustering guna mengevaluasi kinerja mahasiswa, mengidentifikasi variabel-variabel penting yang memengaruhi prediksi kelulusan dengan algoritma Naive Bayes, serta mengukur akurasi model prediksi kelulusan dengan mempertimbangkan hasil segmentasi sebagai fitur tambahan dalam klasifikasi. Latar belakang penelitian didasari oleh pentingnya Latar belakang penelitian ini didasari oleh pentingnya tingkat kelulusan tepat waktu sebagai indikator keberhasilan institusi pendidikan tinggi, yang mencerminkan efektivitas proses pembelajaran dan kesiapan lulusan menghadapi dunia kerja. Penelitian ini menggunakan jenis penelitian eksperimental. Sampel penelitian adalah data mahasiswa yang sudah lulus dalam rentang tahun 2020-2024 yang berjumlah 1.000 data. Teknik analisis data meliputi penerapan algoritma K-Means untuk mengelompokkan mahasiswa berdasarkan data numerik, penentuan jumlah cluster optimal dengan metode Elbow, serta penggunaan algoritma Naive Bayes untuk mengidentifikasi variabel penting dan memprediksi kelulusan mahasiswa. Evaluasi akurasi dilakukan menggunakan Cross validation dan confusion matrix. Hasil Akhir penelitian ini adalah interpretasi dan penarikan kesimpulan dari seluruh proses analisis. Setiap hasil dari tahapan K-Means, dan prediksi dianalisis secara menyeluruh dan saling dikaitkan untuk memberikan gambaran utuh mengenai karakteristik mahasiswa, faktor-faktor utama dalam kelulusan, serta performa sistem prediksi yang dibangun.

Kata Kunci: Data Mining, Evaluasi Akurasi, K-Means, Kelulusan Mahasiswa, Naive Bayes

ABSTRACT

This study aims to group students based on academic data using the K-Means Clustering algorithm to evaluate student performance, identify important variables that influence graduation predictions with the Naive Bayes algorithm, and measure the accuracy of the graduation prediction model by considering the segmentation results as an additional feature in the classification. The background of this study is based on the importance of the on-time graduation rate as an indicator of the success of higher education institutions, which reflects the effectiveness of the learning process and the readiness of graduates to face the world of work. This study uses an experimental research type. The research sample is data from students who have graduated in the period 2020-2024, totaling 1,000 data. Data analysis techniques include the application of the K-Means algorithm to group students based on numerical data, determining the optimal number of clusters with the Elbow method, and the use of the Naive Bayes algorithm to identify important variables and predict student graduation. A accuracy evaluation is carried out using Cross validation and a confusion matrix. The final result of this study is the interpretation and drawing conclusions from the entire analysis process. Each result of the K-Means stage and predictions are analyzed thoroughly and linked to each other to provide a complete picture of student characteristics, key factors in graduation, and the performance of the prediction system built.

Keywords: Data Mining, Accuracy Evaluation, K-Means, Graduation, Naives Bayes

© The Author(s). 2021 Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0>

PENDAHULUAN

Pendidikan tinggi memiliki peran strategis dalam mencetak sumber dayamanusia yang kompeten dan berdaya saing tinggi. Indikator utama keberhasilan institusi pendidikan tinggi adalah tingkat kelulusan mahasiswa yang tepat waktu. Kelulusan tepat waktu tidak hanya mencerminkan efektivitas

proses pembelajaran, tetapi juga berkontribusi pada reputasi institusi dan kesiapan lulusan dalam memasuki dunia kerja.

Universitas Bojonegoro, sebagai institusi pendidikan tinggi, menghadapi tantangan dalam meningkatkan tingkat kelulusan tepat waktu mahasiswanya. Salah satu permasalahan yang diidentifikasi adalah belum dilakukannya analisis mendalam mengenai segmentasi kelulusan mahasiswa berdasarkan target waktu kelulusan. Analisis segmentasi ini penting untuk memahami pola dan faktor yang memengaruhi keterlambatan atau percepatan kelulusan mahasiswa.

Berbagai faktor dapat mempengaruhi kelulusan mahasiswa, yang secara umum dapat dikategorikan menjadi faktor internal dan eksternal. Faktor internal meliputi prestasi akademik, motivasi belajar, dan kemampuan manajemen waktu. Sedangkan faktor eksternal mencakup kualitas pengajaran, kurikulum, serta dukungan fasilitas akademik (Desmawati, 2022). Dalam upaya meningkatkan tingkat kelulusan tepat waktu, penerapan teknologi informasi, khususnya data mining, menjadi relevan. Data mining memungkinkan analisis data dalam jumlah besar untuk menemukan pola dan hubungan yang tidak tampak secara kasat mata. Dalam konteks pendidikan tinggi, teknik ini dapat digunakan untuk memprediksi kelulusan mahasiswa berdasarkan data historis akademik mereka (Suriani, 2023).

Beberapa metode data mining yang umum digunakan dalam prediksi kelulusan mahasiswa antara lain Naive Bayes dan K-Means Clustering. Algoritma Naive Bayes telah diterapkan untuk memprediksi kelulusan mahasiswa dengan mempertimbangkan berbagai variabel akademik. Penelitian oleh Khusnuliawati (2021) mengimplementasikan K-Means untuk mengelompokkan mahasiswa berdasarkan kualitas akademik, yang dapat membantu dalam identifikasi kelompok mahasiswa yang memerlukan perhatian khusus.

Penelitian yang mengambil data yang serupa berupa akademik dan tingkat kelulusan pada mahasiswa dilakukan oleh (Khaerunnisa, 2022) dengan pendekatan metode K-Means untuk menganalisis pola kelulusan mahasiswa berdasarkan IPK dan SKS yang ditempuh menggunakan K-Means Clustering. Penelitian ini mengungkapkan akurasi model Clustering mencapai 96,55%. Penelitian lain yang menggunakan metode Naive Bayes dilakukan oleh (Wahyudi, Wijaya & Khairul, 2024) dengan membandingkan efektivitas K-Means dan Naive Bayes dalam evaluasi tenaga kependidikan berdasarkan SPMI. K-Means hasilkan akurasi 100%, Naive Bayes 91%; K-Means lebih stabil.

Penelitian Rahman (2023) dengan fokus klasifikasi performa akademik siswa yang membandingkan 2 metode yaitu Naive Bayes dan Decision Tree. Hasil dari penelitian ini menyatakan bahwa Naive Bayes unggul (85.97%) dibanding Decision Tree (83.89%); cocok untuk kelompok belajar adaptif. Penggunaan metode yang sama yaitu K-Means dilakukan oleh (Rahman, Mujadilah, Wahyuni & Anas, 2023) dengan mengembangkan model prediksi kelulusan mahasiswa menggunakan K-Means Clustering berdasarkan atribut akademik seperti total SKS, IPS, dan IPK. Metode K-Means Clustering dapat mengelompokkan mahasiswa ke dalam klaster yang berbeda berdasarkan tingkat kelulusan.

Sedangkan Penelitian Jassim, Al-Taie & Naama (2023) mencoba menganalisis dan membandingkan performa mahasiswa dalam pendidikan tradisional vs. E-learning dengan metode data mining. Hasil Clustering menunjukkan 100% tingkat keberhasilan dalam e-learning, dibandingkan dengan maksimal 70% dalam pendidikan tradisional. Penelitian serupa tentang integrasi K-Mean dan Naive Bayes pada kasus yang berbeda untuk klasifikasi wilayah banjir dilakukan oleh (Sinatrya, dkk, 2025) dengan mengklasifikasikan wilayah rawan banjir Jakarta dengan gabungan K-Means dan Naive Bayes. Hasil dari penelitian ini menghasilkan akurasi gabungan mencapai 98.18%, dibanding Naive Bayes konvensional (43.47%).

Dengan menerapkan teknik data mining, seperti algoritma K-Means Clustering, universitas dapat mengelompokkan mahasiswa berdasarkan kinerja akademik dan faktor-faktor lainnya. Hal ini memungkinkan identifikasi kelompok mahasiswa yang berpotensi mengalami keterlambatan kelulusan, sehingga intervensi yang tepat dapat dirancang untuk meningkatkan peluang kelulusantepat waktu. Dengan menerapkan metode prediksi dan segmentasi yang tepat, institusi pendidikan dapat mengidentifikasi mahasiswa yang berpotensi mengalami keterlambatan kelulusan.

Penelitian ini bertujuan untuk mengelompokkan mahasiswa berdasarkan data akademik menggunakan algoritma K-Means Clustering guna mengevaluasi kinerja mahasiswa, mengidentifikasi variabel-variabel penting yang memengaruhi prediksi kelulusan dengan algoritma Naive Bayes, serta mengukur akurasi model prediksi kelulusan dengan mempertimbangkan hasil segmentasi sebagai fitur tambahan dalam klasifikasi.

Berdasarkan latar belakang dan penelitian terdahulu, penulis menggunakan metode Naive Bayes untuk proses klasifikasi prediksi kelulusan mahasiswa dengan terlebih dahulu mengelompokkan variabel pengelompokan mahasiswa yang dilakukan menggunakan metode K-Means pada variabel yang bersifat numerik dengan menentukan jumlah optimal kelompok menggunakan Metode Elbow untuk memperoleh nilai error terkecil. Penggunaan metode cross-validation digunakan untuk proses persebaran data pengujian data testing dan training. Akurasi hasil klasifikasi akan di evaluasi menggunakan confusion matrix untuk mengukur kinerja model dalam memprediksi kelulusan mahasiswa.

METODE PENELITIAN

Penelitian ini merupakan penelitian eksperimental dengan pendekatan deskriptif kuantitatif. Tingkat akurasi model diukur menggunakan confusion matrix untuk menentukan performa terbaik dalam prediksi kelulusan. Objek yang diteliti yaitu data history mahasiswa pada Universitas Bojonegoro berupa data informasi akademik dan history pada tingkat kelulusannya. Data dibagi menjadi data training dan data testing menggunakan metode cross-validation dengan menguji coba data kedalam nilai yang dirubah untuk mencari nilai akurasi yang terbaik. Metode pengumpulan data yang digunakan adalah studi dokumen, dengan mengkaji dataset riwayat akademik mahasiswa Universitas Bojonegoro. Data

yang digunakan merupakan data primer yang diperoleh langsung dari pihak internal universitas, bersifat privat, dan mencakup mahasiswa yang telah lulus pada periode 2020 hingga 2024 sebanyak 1000 data. Teknik analisis data dilakukan terhadap 1.000 data mahasiswa yang telah disesuaikan dalam format .csv agar dapat diolah oleh tools analisis. Langkah pertama dilakukan pengelompokan data mahasiswa yang bersifat numerik menggunakan algoritma K-Means untuk mengetahui jumlah kelompok optimal. Penentuan jumlah cluster terbaik dilakukan dengan metode Elbow, yaitu dengan mengamati selisih error antara centroid dan data. Selanjutnya, dilakukan analisis menggunakan metode Naive Bayes untuk mengidentifikasi variabel-variabel yang memiliki kontribusi terbesar dalam membentuk komponen data. Analisis ini bertujuan untuk mengetahui variabel mana yang paling berpengaruh terhadap label data sebelum dilakukan proses prediksi. Untuk memprediksi kelulusan mahasiswa, digunakan data yang telah dianalisis dengan Naive Bayes, baik berdasarkan seluruh variabel maupun hasil seleksi variabel. Data dibagi menggunakan metode Cross validation untuk menghasilkan data training dan testing. Evaluasi terhadap akurasi model dilakukan dengan menghitung matrik error menggunakan confusion matrix, berdasarkan skenario uji coba dengan parameter terbaik.

HASIL DAN PEMBAHASAN

Kelulusan mahasiswa adalah pencapaian akhir dalam proses pendidikan tinggi yang ditandai dengan penyelesaian seluruh beban akademik yang persyaratan oleh perguruan tinggi, termasuk penyusunan tugas akhir atau skripsi, serta memenuhi standar kompetensi yang ditetapkan (Standar Nasional Pendidikan Tinggi, Permenristekdikti No. 44 Tahun 2015). Berdasarkan regulasi tersebut, syarat kelulusan mahasiswa meliputi penyelesaian sejumlah Satuan Kredit Semester (SKS) sesuai jenjang pendidikan yang ditempuh, pencapaian Indeks Prestasi Kumulatif (IPK) minimal sesuai ketentuan perguruan tinggi, serta pemenuhan persyaratan administratif lainnya seperti publikasi ilmiah atau sertifikasi kompetensi tertentu.

Data mining merupakan kegiatan penemuan pola-pola yang menarik dari data berukuran besar yang disimpan dalam basis data, data warehouse, atau sarana penyimpanan yang lain. Data mining dapat diklasifikasikan menjadi dua kategori: descriptive data mining dan predictive data mining (Hermadi, 2007). Data mining sering disebut sebagai Knowledge Discovery in Database (KDD) yang bertugas untuk mengekstrak pola atau model dari data dengan menggunakan suatu algoritma yang spesifik (Wirdasari & Calam, 2011).

Dari beberapa pengertian tersebut dapat ditarik kesimpulan data mining merupakan suatu teknik menggali informasi berharga yang terpendam atau tersembunyi pada suatu koleksi data (database) yang sangat besar sehingga ditemukan suatu pola yang menarik yang sebelumnya tidak diketahui. Data mining itu sendiri merupakan usaha untuk mendapatkan sedikit barang berharga dari sejumlah besar material dasar oleh karena itu data mining memiliki akar yang panjang dari bidang ilmu

seperti kecerdasan buatan (artificial intelligent), machine learning, statistik dan database (Wirdasari & Calam, 2011). Tahapan untuk melakukan data mining seperti berikut:

Pembersihan data (data cleaning) merupakan proses menghilangkan noise dan data yang tidak konsisten atau yang tidak relevan. Integrasi data (data integration) merupakan penggabungan data dari berbagai database ke dalam satu database baru. Seleksi data (data selection) merupakan data yang tidak sesuai atau yang tidak perlu dianalisis tidak diambil sedangkan data yang sesuai untuk dianalisis yang akan diambil di database. Transformasi data (data transformation) merupakan data yang akan diubah atau digabungkan ke dalam format yang sesuai untuk diproses dalam data mining. Proses mining merupakan suatu proses utama saat metode diterapkan untuk menentukan pengetahuan berharga dan tersembunyi dari data. Evaluasi pola (pattern evaluation) merupakan pengidentifikasian pola-pola menarik ke dalam knowledge based yang ditemukan. Presentasi pengetahuan (knowledge presentation) merupakan visualisasi dan penyajian pengetahuan mengenai metode yang digunakan untuk memperoleh pengetahuan yang diperoleh pengguna. K-Means merupakan algoritma yang sederhana dan cukup efektif dalam proses pengelompokan data berdasarkan kemiripannya atau disebut Clustering.

Pada algoritma K-Means penentuan awal titik centroid didapatkan secara random, dan pada iterasi berikutnya titik centroid didapatkan dari perhitungan jarak rata-rata dari anggota cluster terhadap titik centroid awal. Adapun algoritma K-Means adalah sebagai berikut:

1. Tentukan Nilai K terlebih dahulu. Nilai K adalah jumlah duster yang ingin dibuat.
2. Generate titik centroid awal secara random, apabila $K=2$ maka diperlukan 2 titik centroid awal yang didapatkan secara acak dari dataset.
3. Hitung jarak antara data x_i dengan titik centroid menggunakan rumus Eudidean Distance yang ditampilkan pada Persamaan 1.

$$D((X_i, C_j)) = \sqrt{\sum_{i=1}^p (X_i - C_j)^2}$$

Keterangan=

X_i = data ke-i

C_i = titik pusat duster atau (centroid)

P = dimensi data

4. Menentukan setiap objek data masuk kedalam cluster dengan jarak euclidean terdekat atau terkecil.
5. Melakukan perhitungan ulang titik centroid pada iterasi berikutnya dengan menghitung rata-rata jarak antara semua objek dalam cluster.
6. Ulangi langkah ke-3 hingga nilai centroid tidak berubah atau telah dalam kondisi berhenti (stopping condition).

Metode Elbow merupakan suatu metode yang digunakan untuk menghasilkan informasi dalam menentukan jumlah cluster terbaik dengan caramelihat persentase hasil perbandingan antara jumlah cluster yang akan membentuk siku pada suatu titik.

Pada metode Elbow nilai cluster terbaik yang akan diambil dari nilai Sumof Square Error (SSE) yang mengalami penurunan yang signifikan dan berbentuk siku (Irwanto, 2012). Untuk menghitung SSE menggunakan rumus yang ditunjukkan dalam Persamaan 2.

$$SSE = \sum_{k=1}^K \sum_{x_i \in S_k} \|X_i - C_j\|^2$$

Sum of Square Error (SSE) merupakan rumus yang digunakan untuk mengukur perbedaan antara data yang diperoleh dengan model perkiraan yang telah dilakukan sebelumnya. SSE sering digunakan sebagai acuan penelitian terkait dalam menentukan optimal cluster.

Naive Bayes merupakan algoritma yang digunakan dalam proses klasifikasi yang banyak digunakan dalam pengklasifikasian dokumen, khususnya pada kasus klasifikasi teks. Metode Naive Bayes adalah salah satu metode yang digunakan untuk pengklasifikasian dengan dasar teorema Bayes yang mengasumsikan data bersifat independen dan memiliki arti bahwa data tidak saling terkait satu sama lain (Afriansyah, 2024).

Rumus umum Naive Bayes dapat dilihat dalam persamaan 3.

$$P(C|X) = \frac{P(x|c)P(c)}{P(x)}$$

Keterangan:

- X : Data dengan class yang belum diketahui
- c : Hipotesis data merupakan suatu class spesifik
- $P(c|X)$: Probabilitas hipotesis berdasar kondisi (posteriori probability)
- $P(c)$: Probabilitas hipotesis (prior probability)
- $P(X|c)$: Probabilitas berdasarkan kondisi pada hipotesis
- $P(X)$: Probabilitas c

Cross validation adalah prosedur untuk memperkirakan generalisasi kinerja dalam sebuah metode pemodelan. Cross-validation adalah metode yang paling sering digunakan untuk evaluasi kinerja prediktif dari sebuah model, yakni model yang diberikan sebelumnya atau model yang telah dikembangkan oleh prosedur pemodelan (Yadav & Shukla, 2016).

KE SIMPULAN

Penelitian ini direncanakan untuk menggunakan gabungan algoritma K-Means dan Naive Bayes dalam menganalisis dan memprediksi kelulusan mahasiswa berdasarkan kinerja akademik mahasiswa. K-Means diharapkan dapat mengelompokkan mahasiswa berdasarkan data akademik, sedangkan Naive

Bayes digunakan untuk memperkirakan kemungkinan kelulusan tepat waktu. Melalui pendekatan ini, penelitian diharapkan dapat memberikan kontribusi dalam mendukung strategi evaluasi akademik dan pengambilan keputusan di lingkungan perguruan tinggi

DAFTAR PUSTAKA

- Desmawati, O. (2022). Faktor Yang Menentukan Kelulusan Tepat Waktu Pada Mahasiswa PSPM UIN RIL dengan Analisis Faktor Confirmatori (CFA). Skripsi. UIN Raden Intan Lampung.
- Hermadi Imam, Imas S. Sitanggang, Edward. (2007). Clustering menggunakan Self Organizing Maps (studi kasus: Data PPMB IPB).
- Jassim, A. K. Taie, A. A. Naama, Z. S. (2023). Predicting Students Performance by Using Data Mining Methods. *Journal of Kufa for Mathematics and Computer*, 13(1).
- Khaerunnisa, A. (2022). Analisis Tingkat Kelulusan Mahasiswa di Unisba dengan menggunakan Algoritma K-Means Clustering. *Jurnal Riset Matematika*, 67-76.
- Khusnuliawati, Hardika and Putri, Dhian Riskiana. (2021). Identifikasi Segmen Pasar Mahasiswa Perguruan Tinggi Menggunakan Analisis Klaster Berdasarkan Variabel Psikografis. *Risenologi*, 6 (1b). pp. 44-49. ISSN 2502- 5643.
- Rahman, A. (2023). Klasifikasi Performa Akademik Siswa Menggunakan Metode Decision Tree dan Naive Bayes. *Jurnal Saintekom*, 13 (1).
- Rahman, F. I. Mujadilah, S. Wahyuni, T. Anas, L. (2023). Prediksi Tingkat Kelulusan Menggunakan K-Means Pada Program Studi Informatika Unismuh Makassar. *Jurnal Fasilkom. Universitas Muhammadiyah Makassar*, 13 (3).
- Sinatriya, I. M. Pohan, A. B. Yunita. Amalia, H. Lestari, A. F. Penerapan Integrasi Algoritma K-Means Dan Naive Bayes Untuk Klasifikasi Wilayah Rawan Banjir di Jakarta. *Jurnal Computer Science*, 5 (2).
- Suriani, U. (2023). Penerapan Data Mining untuk Memprediksi Tingkat Kelulusan Mahasiswa Menggunakan Algoritma Decision Tree C4.5. *Journal of Computer and Information Systems Ampera*, 4(2), 55–65.
- Wahyudi, E. Wijaya, R. F. Khairul. (2024). K- Means and Naive Bayes Algorithms for Evaluation of Education Personnel Performance Based on SPMI Standards. *Sinkron*, 8 (3).
- Yadav, S. dan Shukla, S. (2016). Analysis of K-Fold Cross Validation over Hold Out Validation on Colossal Datasets for Quality Classification. *Proceedings 6th International Advanced Computing Conference*, 78-83.