

Membandingkan Teknik Data Mining untuk Memprediksi Prestasi Akademik Mahasiswa

Retno Dwi Handayani¹⁾, Rini Nurlistiani²⁾, Riko Herwanto³⁾

Institut Informatika dan Bisnis Darmajaya

Jl. Z.A Pagar Alam No.93 Labuhan Ratu Bandar Lampung

Telp(0721)787214

e-mail: retnodh84@darmajaya.ac.id¹⁾, rininurlistiani@darmajaya.ac.id²⁾,
rikoherwanto@darmajaya.ac.id³⁾

Abstrak

Tujuan utama perguruan tinggi adalah memberikan pendidikan yang berkualitas kepada para mahasiswanya. Salah satu cara untuk mencapai tingkat kualitas tertinggi adalah dengan mengidentifikasi faktor-faktor yang mempengaruhi kinerja akademik dan kemudian mencoba untuk menyelesaikan kelemahan faktor-faktor tersebut. Tujuan dari penelitian yang diusulkan adalah untuk mendapatkan kinerja akademik mahasiswa dengan cara membandingkan Teknik Teknik Data Mining dalam data yang tersedia (mahasiswa dan catatan perkuliahan). Studi melibatkan sampel 3140 data dalam satu tahun akademik, datanya diambil dan disusun dengan menggunakan paket alat Data Mining (Orange Miniconda). Makalah ini menggunakan reduksi fitur dan teknik klasifikasi untuk mengurangi tingkat kesalahan. Hasil eksperimen menunjukkan signifikan hubungan antara praktek, tugas dalam perkuliahan dan tingkat keberhasilannya. Pada, satu sisi jumlah tugas yang diberikan berdampak negatif pada kinerja akademik perkuliahan. Dalam konteks faktor mempengaruhi kinerja akademik siswa, faktor yang paling mempengaruhi adalah kehadiran siswa di kelas selain ujian akhir dan nilai ujian tengah semester

Kata kunci: Prestasi akademik, Prediksi, Data Mining Pendidikan

1. PENDAHULUAN

Prestasi akademik siswa dipengaruhi oleh banyak faktor, seperti pribadi, sosial ekonomi dan variabel lingkungan lainnya[1]. Pengetahuan tentang faktor-faktor ini dan pengaruhnya terhadap kinerja siswa dapat membantu mengelola pengaruhnya. Baru-baru ini, banyak perhatian telah diberikan pada penelitian Data Mining Pendidikan. Data Mining Pendidikan (*Educational Data Mining*) mengacu pada teknik, alat dan penelitian yang dirancang untuk secara otomatis mengekstraksi makna dari Big Data repository yang dihasilkan oleh atau terkait dengan kegiatan belajar orang di lingkungan pendidikan [2]. Memprediksi kinerja siswa menjadi lebih menantang karena menggunakan database dengan volume data yang besar [3]. Topik penjelasan dan prediksi kinerja akademis banyak diteliti. Kemampuan memprediksi kinerja siswa sangat penting dalam lingkungan pendidikan. Meningkatkan keberhasilan siswa merupakan tujuan jangka panjang di semua institusi akademik. Jika lembaga pendidikan dapat memprediksi kinerja akademik siswa lebih awal sebelum ujian akhir, maka upaya ekstra dapat dilakukan untuk mengatur dukungan yang tepat bagi siswa yang berprestasi rendah untuk meningkatkan studi mereka dan membantu mereka untuk sukses [3]. Mengidentifikasi atribut yang mempengaruhi tingkat keberhasilan perkuliahan dapat membantu peningkatan perkuliahan. Teknologi pendidikan berbasis *web* yang dikembangkan dan penerapan standar kualitas menawarkan kesempatan unik bagi peneliti untuk mempelajari bagaimana siswa belajar dan pendekatan pembelajaran apa yang mengarah pada kesuksesan.

Tujuan utama dari makalah ini adalah untuk mengidentifikasi faktor yang mempengaruhi tingkat keberhasilan mata kuliah dan tingkat keberhasilan siswa kemudian menggunakan faktor-faktor tersebut sebagai prediktor awal untuk tingkat keberhasilan yang diharapkan dan penanganan kelemahannya. Bagian-bagian selanjutnya disusun sebagai berikut: Bagian II berisi apa yang telah dilakukan di bidang *Data Mining* Pendidikan. Bagian III menjelaskan model prediksi yang diusulkan dan kumpulan data yang digunakan. Kemudian bagian berikut menyoroti analisis dan hasil.

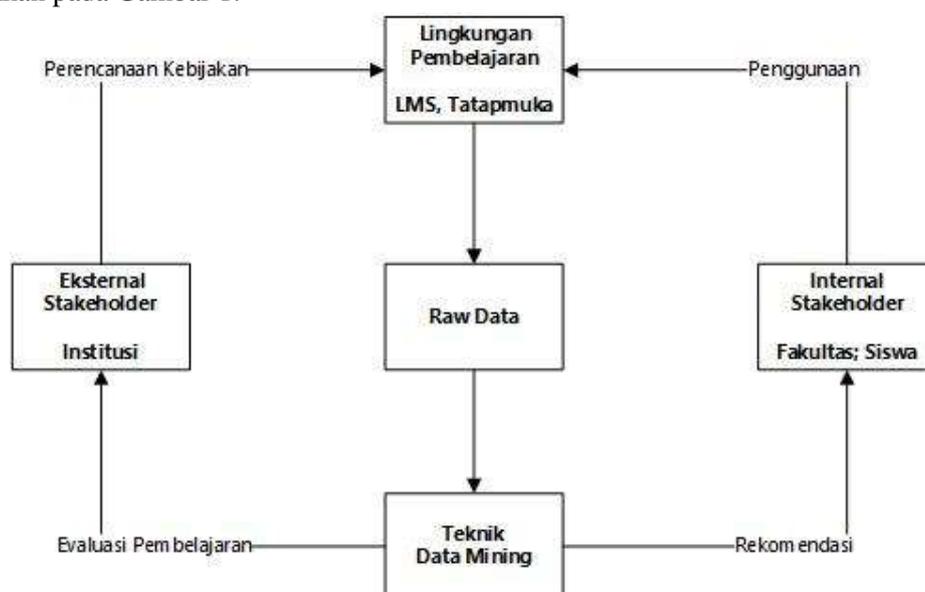
2. METODE PENELITIAN

2.1. Teknik *Data Mining*

Data Mining adalah metode komputasi untuk memproses data berhasil diterapkan di berbagai bidang yang bertujuan untuk memperoleh pengetahuan yang bermanfaat dari data [4]. Teknik *Data Mining* digunakan untuk membangun model untuk mengidentifikasi informasi pengetahuan baru [5]. Ada beberapa teknik *Data Mining* utama yang telah dikembangkan dan digunakan diantaranya asosiasi, klasifikasi, pengelompokan, prediksi, pola sekuensial dan pohon keputusan.

2.2. Model *Data Mining* Pendidikan

Data Mining (DM) mengacu pada penemuan pola yang dapat diasosiasikan dari kumpulan data besar. Ini adalah alat yang ampuh dalam kecerdasan buatan (AI) dan memfasilitasi kategorisasi data ke dalam dimensi yang berbeda, mengidentifikasi hubungan dan mengkategorikan informasi. Hal ini memungkinkan pemangku kepentingan untuk menggunakan informasi yang diambil dari DM dan dapat membantu meningkatkan pengambilan keputusan. *Educational Data Mining* (EDM) adalah disiplin yang berkembang yang digunakan untuk menemukan pengetahuan yang bermakna dan berguna dari data yang diambil dari pengaturan pendidikan. Teknik DM, EDM memberi para peneliti pemahaman yang lebih baik tentang perilaku siswa dan pengaturan di mana pembelajaran terjadi, seperti ditunjukkan pada Gambar 1.



Gambar 1. Siklus penerapan *Data Mining* dalam pendidikan

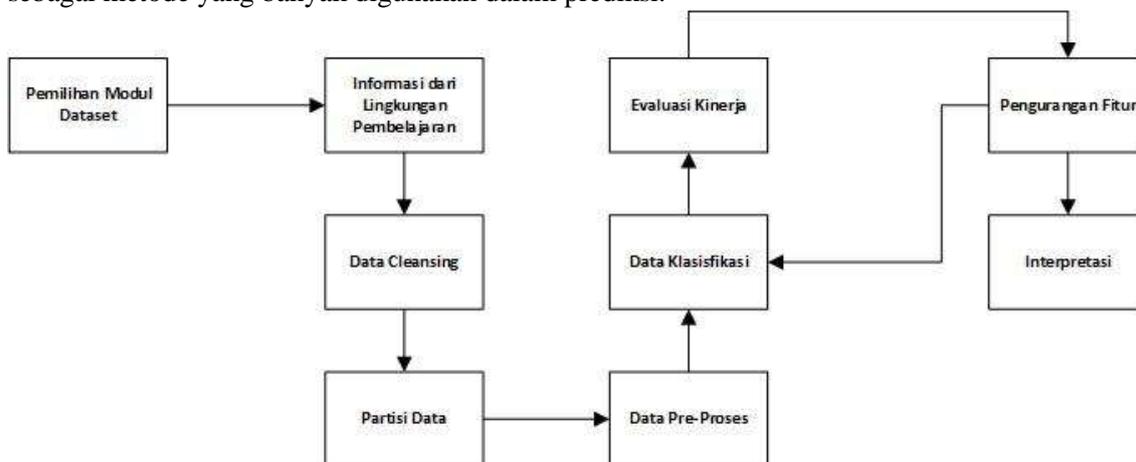
Sumber: *Predicting Student Performance in Higher Educational Institutions
Using Video Learning Analytics and Data Mining Technique*

Proses EDM mengubah data mentah dari berbagai sistem yang digunakan menjadi informasi berguna yang memiliki dampak potensial pada praktik dan penelitian pendidikan. Ini juga disebut sebagai *Knowledge Discovery in Databases (KDD)* karena sifat hierarki dari data yang digunakan [18]. Metode yang digunakan untuk KDD atau EDM adalah sebagai berikut:

- a. *Klasifikasi (Classification)*
Klasifikasi adalah teknik *Data Mining* klasik berdasarkan *Machine Learning*. Pada dasarnya, klasifikasi digunakan untuk mengklasifikasikan setiap item dalam satu set data menjadi salah satu dari sekumpulan kelas atau kelompok yang telah ditentukan sebelumnya. Metode klasifikasi menggunakan teknik matematika seperti pohon keputusan, pemrograman linier, jaringan saraf dan statistik. Dalam klasifikasi, kami mengembangkan perangkat lunak yang dapat mempelajari cara mengklasifikasikan item data menjadi beberapa kelompok terpisah [6]
- b. *Peramalan (Prediction)*
Ini adalah teknik yang banyak digunakan di EDM. Di sini, data historis digunakan dalam hal nilai siswa, data demografis, dll., Untuk memprediksi hasil siswa di masa mendatang. Banyak penelitian telah dilakukan di mana klasifikasi adalah metode yang paling umum digunakan untuk membuat prediksi.
Untuk mencapai target penelitian akan digunakan *Random Forest, Naïve Bayes, Support Vector Machines, Linear Regression* atau model *Logistic Regression*, dan pendekatan K means. Studi ini terutama menyarankan prediksi kinerja akademik siswa baik sebelum, di tengah sesi, atau di akhir semester.
- c. *Pengelompokan (Clustering)*
Clustering adalah teknik *Data Mining* yang membuat cluster objek yang bermakna atau berguna yang memiliki karakteristik serupa menggunakan teknik otomatis. Teknik pengelompokan mendefinisikan kelas dan meletakkan objek di setiap kelas, sedangkan dalam teknik klasifikasi, objek ditugaskan ke dalam kelas yang telah ditentukan. Untuk memperjelas konsepnya, kita bisa mengambil contoh pengelolaan buku di perpustakaan. Dengan menggunakan teknik clustering, kita dapat menyimpan buku-buku yang memiliki beberapa kesamaan dalam satu cluster [6].

Ini adalah proses menemukan dan mengelompokkan sekumpulan objek, yang disebut cluster, dalam grup yang sama berdasarkan sifat yang serupa. Teknik ini digunakan dalam EDM, dimana partisipasi siswa dalam forum online, kelompok diskusi atau obrolan dipelajari. Studi menunjukkan bahwa klasifikasi harus digunakan setelah pengelompokan. Peneliti telah menggunakan algoritme yang serupa dengan algoritme prediksi untuk menentukan keberhasilan akademis siswa.
- d. *Menghubungkan (Associating)*
Metode ini digunakan untuk mencari pola serupa dari beberapa tabel dibandingkan dengan metode lain. Metode yang biasa digunakan untuk menyelidiki hubungan adalah asosiasi, korelasi, dan pola sekuensial. Studi tersebut menyarankan bahwa, dalam EDM, aturan asosiasi harus digunakan untuk memprediksi hasil ujian akhir semester dan kinerja menggunakan algoritma heuristic.

Beberapa aktivitas penting diakui dalam pengkajian data pendidikan, seperti yang ditunjukkan pada Gambar 2. Metode klasifikasi digunakan untuk memprediksi kinerja akademik siswa sebagai metode yang banyak digunakan dalam prediksi.



Gambar 2. Model penerapan penelitian *Data Mining* Pendidikan.

Untuk studi ini, kami juga menggunakan metode klasifikasi, dan data dikumpulkan dari beberapa sistem yang sudah berjalan informasi akademik siswa dikumpulkan dari Sistem Informasi Siswa (SIS), aktivitas online siswa dari LMS Moodle dan data interaksi video siswa dari zoom (aplikasi mobile).

Algoritma yang digunakan untuk penelitian ini didasarkan pada frekuensi. Algoritma yang digunakan berupa literatur yang ada: *Classification Tree*, *Random Forest*, *k-Nearest Neighbours (kNN)*, *Support Vector Machine (SVM)*, *Logistic Regression*, *Naïve Bayes*, *Neural Network* dan *CN2 Induksi Aturan*.

Pemilihan Modul

Penelitian dilakukan dengan kumpulan data mata kuliah berpraktikum dan mata kuliah tidak berpraktikum. dari 3140 siswa yang terdaftar di mata kuliah semester enam.

Data Cleansing

Dalam kegiatan ini, data yang tidak perlu dibersihkan dan data dipisahkan dari informasi yang tidak relevan untuk dianalisis. Data historis untuk setiap siswa yang terdaftar di modul tersebut dipertimbangkan untuk penelitian ini.

Partisi Data

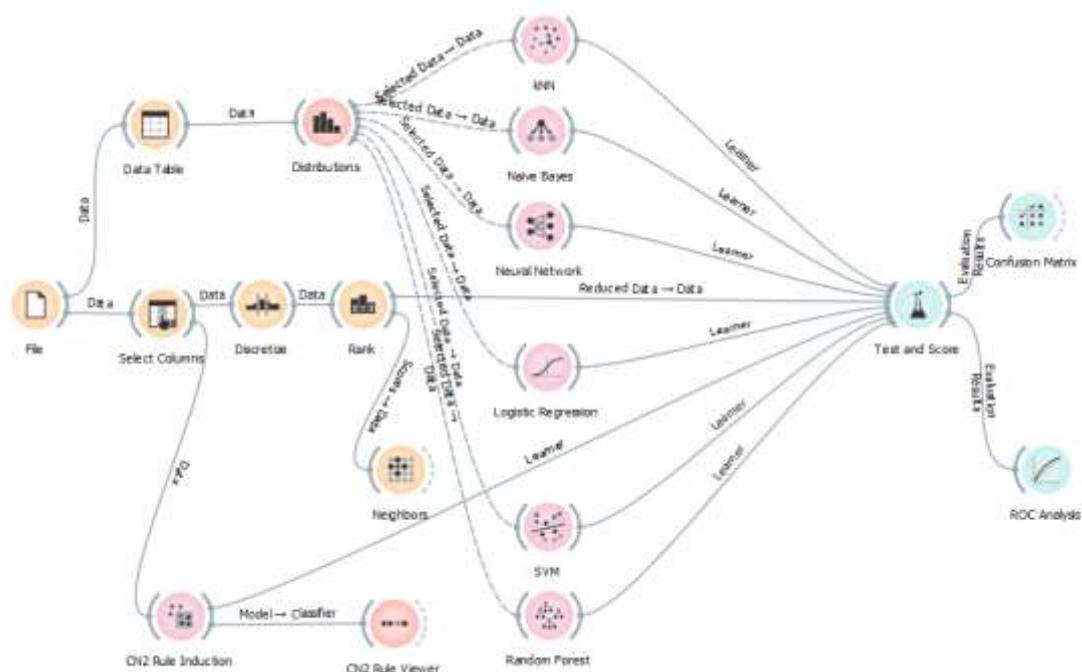
Setelah aktivitas pembersihan data, dilakukan partisi data. Data yang relevan diekstraksi dan dipilah untuk analisis lebih lanjut.

Pra-Pemrosesan Data

Tujuan dilakukannya prapemrosesan data pada sebuah data sebelum melakukan Data Mining adalah sebagai berikut :

1. Untuk mempermudah memahami data sehingga mempermudah pemilihan teknik dan metode data mining yang tepat.
2. Untuk meningkatkan kualitas data sehingga hasil data mining menjadi lebih baik.
3. Untuk meningkatkan efisiensi dan kemudahan poses penambangan data.

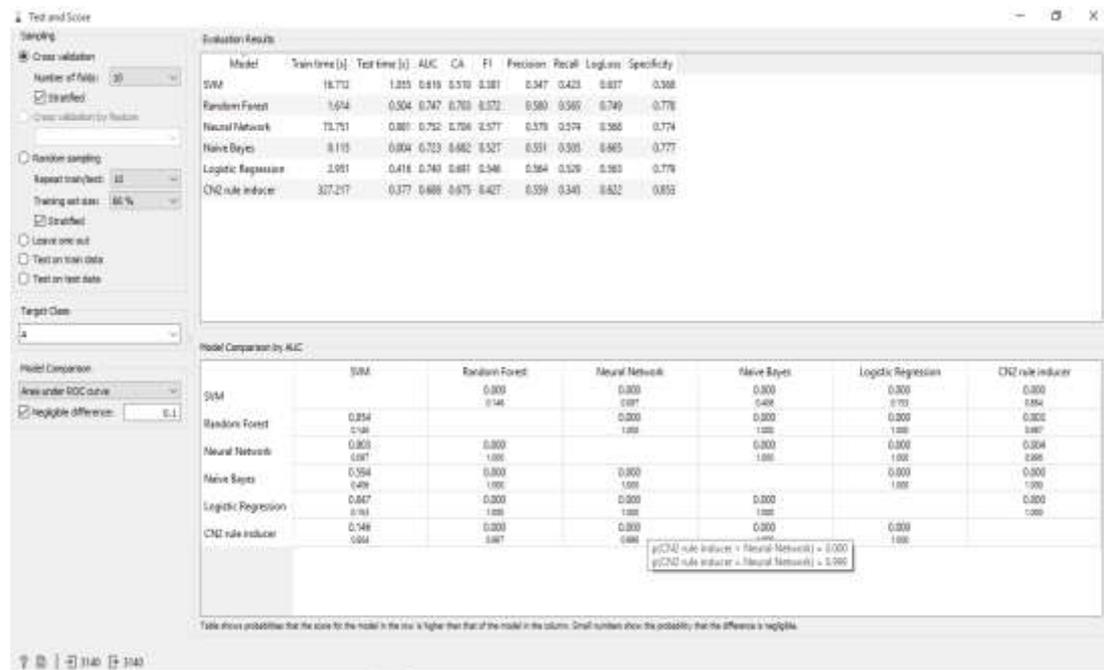
Alasan utama untuk menguji beberapa algoritme pada kumpulan data adalah karena performanya bervariasi untuk fitur yang dipilih. Studi ini menyarankan bahwa algoritma berperilaku berbeda; tergantung pada kumpulan data, efisiensi dan kinerja juga dapat bervariasi. Dengan pendekatan ini, lebih mudah untuk mengidentifikasi salah satu algoritma yang sesuai dengan *dataset* dengan akurasi dan performa yang lebih baik. Untuk tujuan ini, pendekatan serupa digunakan untuk penelitian ini. Untuk studi, alat *Data Mining Orange* digunakan sesuai dengan proses yang ditunjukkan pada Gambar 3.



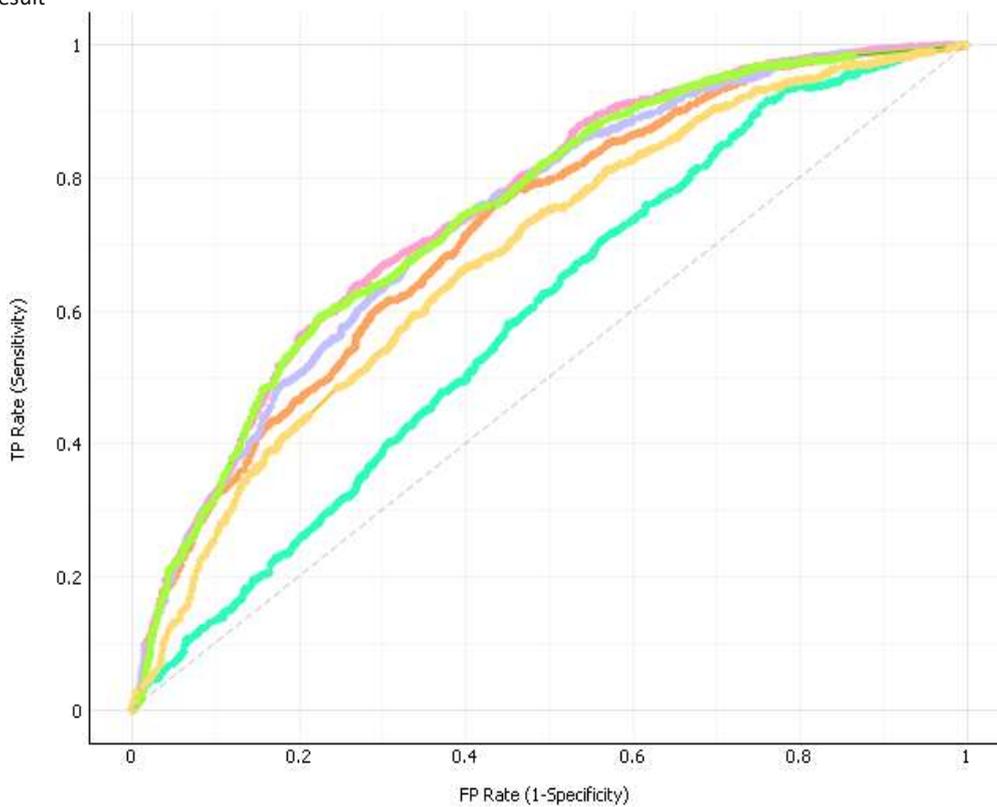
Gambar 3. Model Kerja penelitian Pemrosesan *Data Mining* Pendidikan

3. HASIL DAN PEMBAHASAN

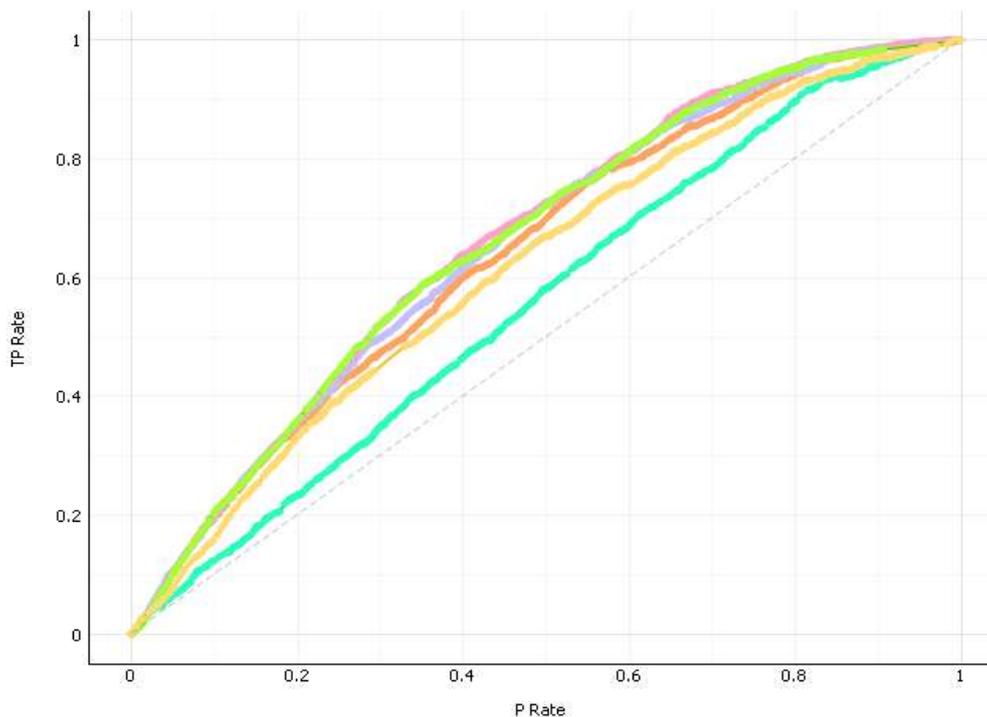
Teknik klasifikasi data terbimbing digunakan untuk menentukan model prediksi terbaik yang sesuai dengan persyaratan untuk memberikan hasil yang optimal. Untuk analisis, sekumpulan algoritma klasifikasi, metrik kinerja dan metode validasi silang 10 kali lipat digunakan.



ROC Result



Lift Curve



Enam analisis dilakukan dengan tujuan untuk menentukan metode dan algoritma pra-pemrosesan mana yang paling baik memprediksi keberhasilan siswa dengan menggunakan seluruh kumpulan data. Dalam analisis pertama, kinerja klasifikasi dihitung tanpa menggunakan data untuk operasi apa pun (model dasar) sedangkan pada analisis kedua dan ketiga, pengaruh teknik transformasi data yang berbeda terhadap kinerja klasifikasi diuji. Dalam tiga analisis berikutnya, efek dari teknik pemilihan fitur yang berbeda pada kinerja klasifikasi diselidiki. Harus dicatat bahwa algoritma klasifikasi yang sama, metrik kinerja, dan metode validasi silang 10 kali lipat digunakan dalam semua analisis. Tidak ada perubahan yang dilakukan pada fitur pada analisis pertama dan semua fitur dimasukkan dalam analisis tanpa melakukan transformasi data apa pun. Algoritma Neural Network mencapai akurasi klasifikasi tertinggi dengan tingkat 83%. Algoritme lain menunjukkan tingkat kinerja 65% ke atas (Table 1).

Table 1. Data Transformasi

Algorithm	Accuracy	Sensitivity	Specificity	F-Measure
Classification Tree	0.65	0.48	0.73	0.49
CN2 Rules	0.66	0.52	0.73	0.52
Naive Bayes	0.74	0.63	0.80	0.63
Neural Network	0.82	0.67	0.90	0.72
kNN	0.76	0.67	0.82	0.67
Random Forest	0.75	0.56	0.86	0.61
SVM	0.80	0.59	0.92	0.68

Analisis hasil yang dilakukan dengan menggunakan fitur yang ditetapkan melalui teknik frekuensi yang sama dan lebar yang sama disajikan masing-masing pada Tabel 2 dan 3.

Table 2. Data Transformation: Equal Frequency, Feature Selection: None

Algorithm	Accuracy	Sensitivity	Specificity	F-Measure
Classification Tree	0.75	0.56	0.86	0.61
CN2 Rules	0.75	0.52	0.88	0.60
Naive Bayes	0.74	0.63	0.80	0.63
Neural Network	0.83	0.74	0.88	0.75
kNN	0.73	0.63	0.78	0.62
Random Forest	0.79	0.63	0.88	0.68
SVM	0.78	0.37	1.00	0.54

Table 3. Data Transformation: Equal Width, Feature Selection: None

Algorithm	Accuracy	Sensitivity	Specificity	F-Measure
Classification Tree	0.79	0.81	0.78	0.73
CN2 Rules	0.77	0.59	0.86	0.64
Naive Bayes	0.75	0.63	0.82	0.64
Neural Network	0.83	0.78	0.86	0.76
kNN	0.77	0.67	0.82	0.67
Random Forest	0.78	0.67	0.84	0.68
SVM	0.75	0.44	0.92	0.56

Table 4. Data Transformation: Equal Width, Feature Selection: Gini Index

Algorithm	Accuracy	Sensitivity	Specificity	F-Measure
Classification Tree	0.79	0.81	0.78	0.73
CN2 Rules	0.86	0.85	0.86	0.81
Naive Bayes	0.76	0.74	0.78	0.69
Neural Network	0.80	0.74	0.84	0.73
kNN	0.86	0.89	0.84	0.81
Random Forest	0.78	0.74	0.80	0.70
SVM	0.78	0.59	0.88	0.65

Ketika kinerja dari metode transformasi data dibandingkan dengan model dasar (Tabel 4), diamati bahwa algoritma meningkatkan kinerja klasifikasinya dibandingkan dengan model berbasis di kedua metode transformasi data (dengan pengecualian algoritma SVM). Penggunaan algoritma berbasis pohon (*Classification Tree dan CN2 Rules*) meningkatkan akurasi klasifikasi dengan kisaran 9% hingga 14% melalui konversi data ke dalam bentuk kategorikal. Ketika membandingkan analisis yang dilakukan dengan menggunakan frekuensi yang sama dan ukuran lebar yang sama, teknik lebar yang sama menunjukkan kinerja yang lebih baik dibandingkan dengan semua algoritma dalam memprediksi siswa yang gagal (berdasarkan metrik sensitivitas).

4. SIMPULAN

Memprediksi kinerja mahasiswa sebagian besar berguna untuk membantu pendidik dan peserta didik meningkatkan pembelajaran mereka dan proses pengajaran. Makalah ini mengulas penelitian sebelumnya tentang memprediksi kinerja mahasiswa dengan berbagai analitik metode. Sebagian besar peneliti telah menggunakan rata-rata nilai kumulatif (CGPA) dan

penilaian internal sebagai *data set*. Sedangkan untuk teknik prediksi, metode klasifikasi sering digunakan di bidang pendidikan data mining. Di bawah teknik klasifikasi, Neural Network dan Decision Tree adalah dua metode baik dipergunakan untuk memprediksi kinerja mahasiswa. Kesimpulannya, meta-analisis memprediksi kinerja mahasiswa akan memotivasi untuk melakukan penelitian lebih lanjut untuk diterapkan di lingkungan ini akan membantu sistem pendidikan untuk memantau kinerja mahasiswa secara sistematis.

DAFTAR PUSTAKA

- [1] Sana, B., Siddiqui, I.F. and Arain, Q.A., 2019. Analyzing students' academic performance through educational data mining.
- [2] Nithya P, Umamaheswari B, Umadevi A., 2016.) A survey on educational Data Mining in field of education. *J Comput Sci Softw Dev* 1: 1-6.
- [3] Rachana, G., Karnik, S. R. R., Singh, S., & Lokesh, S., 2019. Academic Performance Analysis System. *International Research Journal of Engineering and Technology*, 6(4), 5169-5172.
- [4] Azevedo, A., 2019. Data Mining and knowledge discovery in databases. In *Advanced Methodologies and Technologies in Network Architecture, Mobile Computing, and Data Analytics* (pp. 502-514). IGI Global.
- [5] Yaacob, W. F. W., Nasir, S. A. M., Yaacob, W. F. W., & Sobri, N. M., 2019. Supervised Data Mining approach for predicting student performance. *Indones. J. Electr. Eng. Comput. Sci*, 16, 1584-1592.
- [6] [Verma K, Singh A, Verma P., 2016. A review on predicting students performance using Data Mining techniques. *IJCESR* 3: 127-132.
- [7] [Altaher A, BaRukab O., 2017. Prediction of student's academic performance based on adaptive neuro-fuzzy inference. *IJCSNS* 17: 165-169.
- [8] Sekeroglu, B., Dimililer, K., & Tuncal, K., 2019. Student performance prediction and classification using machine learning algorithms. In *Proceedings of the 2019 8th International Conference on Educational and Information Technology* (pp. 7-11).
- [9] Agaoglu M., 2016. Predicting instructor performance using data mining techniques in higher education.
- [10] Ramesh V, Parkavi P, Ramar K (2016) Predicting student performance: A statistical and Data Mining approach. *IJCA* 63: 35-39.
- [11] Salal, Y. K., Abdullaev, S. M., & Kumar, M., 2019. Educational Data Mining: Student Performance Prediction in Academic. *IJ of Engineering and Advanced Tech*, 8(4C), 54-59.
- [12] Sultana, J., Usha, M., & Farquad, M. A. H., 2019. Student's Performance Prediction using Deep Learning and Data Mining Methods. *International Journal of Recent Technology and Engineering (IJRTE)/Blue Eyes Intelligence Engineering & Sciences Publication*, (1S4), 1018-1021.
- [13] Romero, C., & Ventura, S., 2020. Educational Data Mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3), e1355.
- [14] Hasan, R.; Palaniappan, S.; Mahmood, S.; Naidu, V.R.; Agarwal, A.; Singh, B.; Sarker, K.U.; Abbas, A.; Sattar, M.U. A review: Emerging trends of big data in higher educational institutions. In *Micro-Electronics and Telecommunication Engineering; Lecture Notes in Networks and Systems; Springer: Singapore, 2020; Volume 106, pp. 289-297*
- [15] Chaudhury, P.; Tripaty, H.K. An empirical study on attribute selection of student performance prediction model. *Int. J. Learn. Technol.* 2017, 12, 241-252
- [16] Hasan, R.; Palaniappan, S.; Raziff, A.R.A.; Mahmood, S.; Sarker, K.U. Student academic performance prediction by using decision tree algorithm. In *Proceedings of the*

- 2018 4th International Conference on Computer and Information Sciences: Revolutionising Digital Landscape for Sustainable Smart Society, ICCOINS 2018—Proceedings, IEEE, Kuala Lumpur, Malaysia, 13–14 August 2018; pp. 1–5.
- [17] Akçapınar, G.; Altun, A.; Askar, P. Using learning analytics to develop early-warning system for at-risk students. *Int. J. Educ. Technol. High. Educ.* 2019, 16, 1–20.
- [18] Saqr, M.; Fors, U.; Tedre, M. How learning analytics can early predict under-achieving students in a blended medical education course. *Med. Teach.* 2017, 15, 1–11.
- [19] Abu Zohair, L.M. Prediction of Student's performance by modelling small dataset size. *Int. J. Educ. Technol. High. Educ.* 2019, 16
- [20] Yassein, N.A.; M Helali, R.G.; Mohomad, S.B. Predicting student academic performance in KSA using Data Mining techniques. *J. Inf. Technol. Softw. Eng.* 2017, 7, 1–15