Procedia of Sciences and Humanities
Proceedings of the 1st SENARA 2022

Educational Data Mining on Student Academic Performance Prediction: A Survey

Educational Data Mining Pada Prediksi Kinerja Akademik Mahasiswa : Sebuah Survey

Uce Indahyanti[1]*, Nuril Lutvi Azizah[2], Hamzah Setiawan[3]
[1 23]Fakultas Sains dan Teknologi, Universitas Muhammadiyah Sidoarjo, Indonesia

Abstract. *Student academic performance prediction has become a hot research topic, and is still a research trend in the field of educational data mining (EDM). The application of data mining in the education domain can find some hidden knowledge and patterns, which help in decision making for management to improve the education system. This study presents survey results in the form of systematic mapping of literature related to EDM, which aims to identify methods, datasets, and results obtained by researchers in the last five years.*

Keywords: *educational data mining, student performance prediction, surveys, literature mapping, systematic.*

Abstrak. Prediksi kinerja akademik mahasiswa telah menjadi topik penelitian yang hangat, dan masih menjadi tren riset pada ranah penambangan data pendidikan atau educational data mining (EDM). Penerapan data mining pada domain pendidikan dapat menemukan beberapa pengetahuan dan pola tersembunyi, yang sangat membantu dalam pengambilan keputusan bagi manajemen untuk meningkatkan sistem pendidikan. Penelitian ini menyajikan hasil survey berupa pemetaan literatur terkait EDM secara sistematis, yang bertujuan untuk mengidentifikasi metode, dataset, dan hasil yang diperoleh para peneliti dalam lima tahun terakhir.

Kata kunci: educational data mining, prediksi kinerja mahasiswa, survey, pemetaaan literatur, sistematis.

ISSN 2722-0672 (online), https://pssh.umsida.ac.id. Published by Universitas Muhammadiyah Sidoarjo

1 Pendahuluan

Prediksi kinerja akademik mahasiswa merupakan salah satu implementasi Educational Data Mining (EDM), yang muncul sebagai area baru penelitian karena perluasan berbagai metode statistik yang digunakan untuk pengaturan data pendidikan. Penerapan teknik penambangan data atau data mining pada domain pendidikan dapat menemukan beberapa pengetahuan dan pola tersembunyi, yang akan membantu dalam pengambilan keputusan bagi manajemen untuk meningkatkan sistem pendidikan. Pada sebuah sistem pendidikan berbasis web, fitur perilaku peserta didik sangat signifikan dalam menunjukkan interaksi antara mahasiswa dengan sistem elearning [1].

Sejumlah penelitian yang berfokus pada prediksi kinerja mahasiswa telah banyak dilakukan, baik dalam pembelajaran berbasis elearning (learning management system) maupun pembelajaran tradisional. Teknik data mining menggunakan pendekatan machine learning maupun statistik telah banyak digunakan untuk menganalisis dan memprediksi kinerja akademik mahasiswa, termasuk menemukan fitur atau atribut yang mempengaruhi hasil studi [2]. Teknik stratified random sampling dan data mining digunakan untuk mendeteksi sedini mungkin resiko kegagalan mahasiswa dengan memperhatikan durasi awal mulainya course [3]. Teknik supervised learning digunakan untuk memprediksi tiga atribut log aktifitas elearning yaitu student enrolments, users, dan courses [4]. Klasifikasi log aktifitas dari beberapa course elearning menggunakan algoritma klasifikasi J48 yang menghasilkan model prediksi dengan pendekatan decision tree [5].

Penelitian lain yang bertujuan untuk meningkatkan kestabilan dan akurasi prediktif menggunakan metode ensemble Random Forest (RF). Kasus yang digunakan dalam penelitian ini adalah klasifikasi ketidaktepatan waktu kelulusan mahasiswa Universitas Terbuka. Dataset penelitian tersebut menggunakan data sekunder berupa data wisuda, yang terdiri dari variable IPK, jurusan/prodi, pendidikan dan pekerjaan orang tua, dan status mahasiswa. Hasil analisis menunjukkan bahwa ensemble RF mampu meningkatkan akurasi klasifikasi ketidaktepatan waktu kelulusan mahasiswa yang mencapai konvergen dengan prediksi klasifikasi mencapai 93.23% [6].

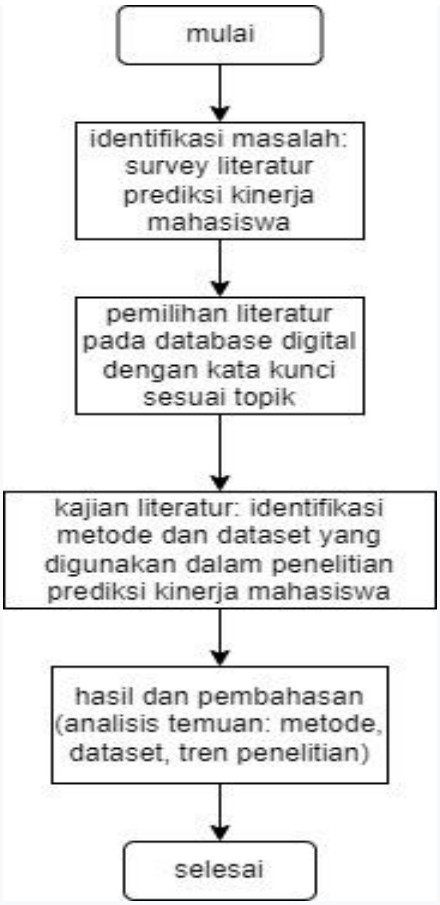
Sebuah literatur review juga menyebutkan bahwa topik terkait prediksi kinerja akademik merupakan topik yang paling banyak diteliti dalam domain EDM. Selain itu topik lain dalam domain EDM yang juga banyak diteliti antara lain decision support for teacher and students, detection of behavioral patterns, comparison or optimization of algorithms, learner modelling, dan predictive analysis of dropout [7].

Pada umumnya metode atau algoritma yang digunakan pada penelitian prediksi kinerja mahasiswa berbasis machine learning antara lain Decision Tree (DT), Support Vector Machine SVM, Multilayer Perceptron (MLP), Convolutional Neural Network (CNN), Logistic Regression (LR), Naïve Bayes (NB), Deep Learning, dan ensemble learning. Sedangkan dataset yang digunakan cukup beragam, antara lain data log elearning (public dan private), data akademik, dan data wisuda. Penelitian ini bertujuan untuk mengidentifikasi dan menganalisis tren penelitian, metode dan dataset yang digunakan dalam penelitian prediksi kinerja siswa lima tahun terakhir yaitu tahun 2017 sampai dengan awal 2022.

2 Metode Penelitian

Penelitian ini menggunakan pendekatan kajian literatur sistematik, dimulai dari identifikasi masalah, pemilihan literatur prediksi kinerja mahasiswa yang dipublikasikan selama lima tahun terakhir, mengidentifikasi metode dan dataset yang digunakan, serta, serta menganalisis temuannya. Diagram alir penelitian ditampilkan dalam Gambar 1. Tahapan Penelitian.

ISSN 2722-0672 (online), https://pssh.umsida.ac.id. Published by Universitas Muhammadiyah Sidoarjo



Gambar 1. Tahapan Penelitian

2.1. Strategi Survey

Pencarian dan pemilihan literatur atau artikel dilakukan mengunakan kata kunci "student performance prediction" dari berbagai database digital antara lain IEEExplore (https://ieeexplore.ieee.org/), Springer Link (https://link.springer.com/), Science Direct (https://www.sciencedirect.com/), dan google scholar (https://scholar.google.com/). Pencarian dibatasi pada tahun publikasi 2017 sampai dengan awal 2022 yang diterbitkan pada jurnal atau prosiding nasional dan internasional bereputasi.

2.2. Pemilihan Literatur

Aplikasi Mendeley digunakan untuk menyimpan dan mengelola hasil pencarian artikel. Sebanyak dua puluh empat literatur atau artikel diperoleh untuk dikaji dan dianalisis sesuai tujuan penelitian ini.

3 Hasil dan Pembahasan

Setelah literatur-literatur diperoleh, selanjutnya dilakukan pemetaan literatur menggunakan tabel seperti yang ditampilkan pada Tabel 1. Telaah Literatur Prediksi Kinerja Mahasiswa di bawah ini, yang dibagi dalam kolom-kolom penulis, dataset, pendekatan dan metode.

Tabel 1. Telaah Literatur Prediksi Kinerja Mahasiswa

Penulis	Dataset	Pendekatan & Metode
		Single classifier:
Kondo et al, 2017 [8]	Data log elearning	Naïve Bayesian, Decision Tree,
Conijn et al, 2017 [9]	Data akademik dan log elearning	K-Nearest Neighbor, Fuzzy K-Nearest Neighbor, Support Vector
Digna et al, 2017 [10]	Data log elearning	Machine, Logistic Regression,
Kaur et al, 2018 [11]	Data akademik	Multi-layer Perceptron, Deep
Son et al, 2019 [12]	Data akademik	Learning (Deep Neural Network),

ISSN 2722-0672 (online), https://pssh.umsida.ac.id. Published by Universitas Muhammadiyah Sidoarjo

Hassan et al, 2019 [13] Olive et al, 2020 [4] Zambrano et al, 2020 [5] Benediktus et al, 2020 [14] Gonzalez et al, 2020 [3] Hashim et al, 2020 [15] Hidalgo et al, 2021 [16] Mubarak et al, 2022 [17]	Data log elearning Data log elearning Data log elearning Data log elearning Data log elearning Data akademik Data log elearning Data log elearning	Deep Learning on Graph (Graph Convolutional Networks)
Abubakar et al, 2017 [18] Salini et al, 2018 [19] Kumari et al, 2018 [20] Suwardika et al, 2019 [6] Arrahimi et al, 2019 [21] Injadat et al, 2020 [2] Ajibade et al, 2020 [1] Verma et al, 2021 [22] Asselman et al, 2021 [23]	Data log elearning Data akademik Data akademik Data wisudawan Data akademik Data akademik Data akademik dan log elearning Data log elearning Data akademik dan log elearning	**Ensemble learning**: Random Forest, AdaBoost, XGBoost, Stacking ensemble
Shreem et al, 2022 [24]	Data akademik	**Hybrid machine learning** approach (Genetic algorithm and five different classifiers)

Analisis temuan terhadap hasil telaah literatur prediksi kinerja mahasiswa di atas, dapat dibagi dalam dua sub bahasan di bawah ini.

3.1. Metode Yang Banyak Digunakan

Berdasarkan kajian literatur di atas, penelitian prediksi kinerja mahasiswa pada umumnya menggunakan algoritma klasifikasi machine learning, baik dengan pendekatan pengklasifikasi tunggal (model dasar) maupun ensemble learning (menggabungkan beberapa model dasar). Machine learning banyak digunakan karena dapat menggantikan atau menirukan perilaku manusia dalam pembuatan keputusan tertentu.

Pengklasifikasi tunggal atau model dasar yang digunakan dalam penelitian di atas, antara lain Naïve Bayesian, Decision Tree, K-Nearest Neighbor, Support Vector Machine, Logistic Regression, dan Deep Learning (Neural Network). Deep learning merupakan subbidang machine learning yang algoritmanya terinspirasi dari struktur otak manusia (jaringan syaraf tiruan). Pada umumnya para peneliti menggunakan lebih dari satu model dasar dalam penelitian mereka yang kemudian dibandingkan masing-masing hasil prediksinya (bukan digabungkan). Disamping itu terdapat studi yang menggunakan teknik hybrid memadukan algoritma genetika yaitu sebuah teknik optimasi dengan algoritma machine learning, kemudian membandingkan hasil keduanya [24].

Ensemble learning merupakan algoritma machine learning yang menggabungkan beberapa model dasar (pengklasifikasi tunggal) untuk memprediksi suatu hasil, dengan teknik majority atau average vote, yang bertujuan untuk mengurangi kesalahan prediksi. Ensemble learning terdiri dari teknik bagging, boosting, dan stacking yang menggabungkan berbagai hasil prediksi dari masing-masing pengklasifikasi menjadi sebuah prediksi akhir. Salah satu teknik penggabungan yang banyak digunakan adalah majority vote. Beberapa penelitian menyatakan ensemble learning mampu menghasilkan prediksi yang lebih akurat dibandingkan model dasar atau tunggal [2] [1] [18].

3.2. Dataset Yang Banyak Digunakan

Dataset penelitian yang digunakan pada umumnya terdiri dari data log aktifitas elearning (diunduh dari LMS) dan non elearning (data akademik). Pada kedua jenis data tersebut mengandung karakteristik utama mahasiswa seperti data demografi, latar belakang akadekik, dan fitur perilaku mahasiswa yang mewakili dataset machine learning [15].

ISSN 2722-0672 (online), https://pssh.umsida.ac.id. Published by Universitas Muhammadiyah Sidoarjo

Dataset penelitian ada yang diambil atau diunduh dari database publik seperti Kaggle Data Science Community (https://www.kaggle.com/) dan UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets.php) berbasis log aktifitas elearning, ada juga yang diambil dari database privat berupa data akademik dan atau log aktifitas elearning perguruan tinggi tertentu. Selain itu terdapat beberapa penelitian yang menggunakan Data log aktifitas elearning yang digunakan pada umumnya berupa data public berplatform Moodle. Moodle merupakan learning management system (LMS) paling populer di dunia yang bersifat free-open source (https://moodle.org/).

4 Kesimpulan

Berdasarkan hasil survey di atas, dapat disimpulkan bahwa penerapan teknik penambangan data atau data mining pada domain pendidikan telah banyak dilakukan dengan berbagai metode dan dataset. Pendekatan machine learning masih menjadi tren dalam penelitian berbasis prediksi termasuk prediksi kinerja mahasiswa, baik menggunakan pengklasifikasi tunggal maupun ensemble learning yang merupakan gabungan dari beberapa pengklasifikasi tunggal. Penelitian lain mengusulkan teknik hybrid yaitu menggunakan algoritma genetika dan machine learning kemudian membandingkan hasil keduanya. Metode ensemble learning dinyatakan mampu menghasilkan prediksi yang lebih akurat dibandingkan pendekatan pengklasifikasi tunggal [2] [1] [18], dan masih terdapat peluang untuk mengembangkannya. Sedangkan dataset penelitian pada umumnya terdapat dua jenis yaitu berbasis log aktifitas elearning dan non elearning yang diunduh dari database publik maupun privat.

Penelitian ini masih mempunyai keterbatasan, saran pengembangan ke depan menambah jumlah literatur khususnya yang dipublikasikan pada tahun 2022, dan mengkaji secara rinci atribut-atribut yang dikandung dalam dataset penelitian.

5 Ucapan terima kasih

Terima kasih kepada Direktorat Riset dan Pengabdian Masyarakat Universitas Muhammadiyah Sidoarjo yang telah memberikan hibah riset internal ini.

Referensi

[1] S. M. S. Samuel-Soma M. Ajibade, Nor Bahiah Ahmad, *A Data Mining Approach to Predict Academic Performance of Students Using Ensemble Techniques*. Springer International Publishing, 2020.

[2] M. N. Injadat, A. Moubayed, A. B. Nassif, and A. Shami, "Systematic ensemble model selection approach for educational data mining," *Knowledge-Based Syst.*, vol. 200, p. 105992, 2020, doi: 10.1016/j.knosys.2020.105992.

[3] M. Riestra-González, M. del P. Paule-Ruíz, and F. Ortin, "Massive LMS log data analysis for the early prediction of course-agnostic student performance," *Comput. Educ.*, vol. 163, no. December 2020, 2021, doi: 10.1016/j.compedu.2020.104108.

[4] D. Monllaó Olivé, D. Q. Huynh, M. Reynolds, M. Dougiamas, and D. Wiese, "A supervised learning framework: using assessment to identify students at risk of dropping out of a MOOC," *J. Comput. High. Educ.*, vol. 32, no. 1, pp. 9–26, 2020, doi: 10.1007/s12528-019-09230-1.

[5] J. López-Zambrano, J. A. Lara, and C. Romero, "Towards portability of models for predicting students' final performance in university courses starting from moodle logs," *Appl. Sci.*, vol. 10, no. 1, 2020, doi: 10.3390/app10010354.

[6] G. S. Suwardika and I. K. P. Suniantara, "Analisis Random Forest Pada Klasifikasi Cart Ketidaktepatan Waktu Kelulusan Mahasiswa Universitas Terbuka," *BAREKENG J. Ilmu Mat. dan Terap.*, vol. 13, no. 3, pp. 177–184, 2019, doi: 10.30598/barekengvol13iss3pp177-184ar910.

[7] X. Du, J. Yang, J. L. Hung, and B. Shelton, "Educational data mining: a systematic review of research and emerging trends," *Inf. Discov. Deliv.*, vol. 48, no. 4, pp. 225–236, 2020, doi: 10.1108/IDD-09-2019-0070.

[8] N. Kondo, M. Okubo, and T. Hatanaka, "Early Detection of At-Risk Students Using Machine Learning Based on LMS Log Data," *Proc. - 2017 6th IIAI Int. Congr. Adv. Appl. Informatics, IIAI-AAI 2017*, pp. 198–201, 2017, doi: 10.1109/IIAI-AAI.2017.51.

[9] R. Conijn, C. Snijders, A. Kleingeld, and U. Matzat, "Predicting student performance from LMS data: A comparison of 17 blended courses using moodle LMS," *IEEE Trans. Learn. Technol.*, vol. 10, no. 1, pp. 17–29, 2017, doi: 10.1109/TLT.2016.2616312.

[10] E. Digna S, "LEARNING MANAGEMENT SYSTEM WITH PREDICTION MODEL AND COURSE-CONTENT RECOMMENDATION MODULE," *Learn. Manag. Syst. WITH Predict. Model COURSE-CONTENT Recomm. Modul.*, vol. 16, no. 1, pp. 437–457, 2017, doi: https://doi.org/10.28945/3883.

ISSN 2722-0672 (online), https://pssh.umsida.ac.id. Published by Universitas Muhammadiyah Sidoarjo

[11] B. S. Amandeep Kaur, Nitin Umesh, "Machine Learning Approach to Predict Student Academic Performance Amandeep," *Int. J. Res. Appl. Sci. Eng. Technol.*, 2018, doi: http://doi.org/10.22214/ijraset.2018.4125.
[12] L. H. Son and H. Fujita, "Neural-fuzzy with representative sets for prediction of student performance," *Appl. Intell.*, vol. 49, no. 1, pp. 172–187, 2019, doi: 10.1007/s10489-018-1262-7.
[13] H. Hassan, S. Anuar, and N. B. Ahmad, *Students' performance prediction model using meta-classifier approach*, vol. 1000. Springer International Publishing, 2019.
[14] N. Benediktus and R. S. Oetama, "Algoritma Klasifikasi Decision Tree C5 . 0 untuk Memprediksi Performa Akademik Siswa," *Ultimatics*, vol. XII, no. 1, pp. 14–19, 2020.
[15] A. S. Hashim, W. A. Awadh, and A. K. Hamoud, "Student Performance Prediction Model based on Supervised Machine Learning Algorithms," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 928, no. 3, 2020, doi: 10.1088/1757-899X/928/3/032019.
[16] Á. C. Hidalgo, P. M. Ger, and L. D. L. F. Valentín, "Using Meta-Learning to predict student performance in virtual learning environments," *Appl. Intell.*, vol. 52, no. 3, pp. 3352–3365, 2022, doi: 10.1007/s10489-021-02613-x.
[17] A. A. Mubarak, H. Cao, I. M. Hezam, and F. Hao, "Modeling students' performance using graph convolutional networks," *Complex Intell. Syst.*, 2022, doi: 10.1007/s40747-022-00647-3.
[18] Y. Abubakar, N. Bahiah, and H. Ahmad, "Prediction of Students' Performance in E-Learning Environment Using Random Forest," *Int. J. Innov. Comput.*, vol. 7, no. 2, pp. 1–5, 2017, [Online]. Available: http://se.fsksm.utm.my/ijic/index.php/ijic.
[19] A. Salini, U. Jeyapriya, S. M. College, and S. M. College, "A Majority Vote Based Ensemble Classifier for Predicting Students Academic Performance," *Int. J. Pure Appl. Math.*, vol. 118, no. 24, pp. 1–11, 2018.
[20] P. Kumari, P. K. Jain, and R. Pamula, "An efficient use of ensemble methods to predict students academic performance," *Proc. 4th IEEE Int. Conf. Recent Adv. Inf. Technol. RAIT 2018*, pp. 1–6, 2018, doi: 10.1109/RAIT.2018.8389056.
[21] A. R. Arrahimi, M. K. Ihsan, D. Kartini, M. R. Faisal, and F. Indriani, "Teknik Bagging Dan Boosting Pada Algoritma CART Untuk Klasifikasi Masa Studi Mahasiswa," *J. Sains dan Inform.*, vol. 5, no. 1, pp. 21–30, 2019, doi: 10.34128/jsi.v5i1.171.
[22] B. K. Verma and H. K. Singh, "Prediction of Students ' Performance in e - Learning Environment using Data Mining / Machine Learning Techniques," vol. 23, no. 5, pp. 586–593, 2021.
[23] A. Asselman, M. Khaldi, and S. Aammou, "Enhancing the prediction of student performance based on the machine learning XGBoost algorithm," *Interact. Learn. Environ.*, vol. 0, no. 0, pp. 1–20, 2021, doi: 10.1080/10494820.2021.1928235.
[24] S. S. Shreem, H. Turabieh, S. Al Azwari, and F. Baothman, "Enhanced binary genetic algorithm as a feature selection to predict student performance," *Soft Comput.*, vol. 26, no. 4, pp. 1811–1823, 2022, doi: 10.1007/s00500-021-06424-7.

ISSN 2722-0672 (online), https://pssh.umsida.ac.id. Published by Universitas Muhammadiyah Sidoarjo