

Prediksi Kelulusan Siswa pada Mata Pelajaran Matematika menggunakan *Educational Data Mining*

Selly Anastassia Amellia Kharis^{1,a)}, Arman Haqqi Anna Zili², Edward Zubir¹, Fauzan Ihza Fajar¹

¹Universitas Terbuka

²Universitas Indonesia

Email: ^{a)} selly@ecampus.ut.ac.id

Abstrak

Pesatnya perkembangan teknologi menyebabkan keterlimpahan data di berbagai bidang termasuk pendidikan. Di dunia pendidikan tentu banyak data yang dapat digali, mulai dari data kinerja pengajar, data nilai mahasiswa, *tracer study*, dan data proses pembelajaran. Di sisi lain, evaluasi pembelajaran yang dilakukan biasanya terlaksana setelah nilai ujian keluar dan tidak ada peringatan dini untuk siswa terkait dengan kemajuan pembelajarannya. Akibatnya, tindakan pencegahan yang dilakukan oleh guru atau pihak sekolah menjadi terlambat dan dapat berdampak hingga ketidaklulusan siswa. Salah satu inovasi evaluasi pembelajaran yang dapat dilakukan di sekolah adalah dengan menggunakan *educational data mining* untuk memprediksi kelulusan siswa. *Educational data mining* diawali dengan pengumpulan, transformasi hingga analisis data. Pada penelitian ini, dilakukan perbandingan berbagai metode *machine learning* untuk memprediksi kelulusan siswa. Dalam penggalian data, penelitian ini menggunakan data akademik siswa menengah pada mata pelajaran matematika di Portugal dengan 395 sampel dan 33 variabel bebas. Hasil dari *educational data mining* diharapkan dapat menjadi rujukan untuk menentukan strategi yang tepat untuk menurunkan ketidaklulusan siswa dan menjadi peringatan dini untuk ketidaklulusan siswa. Berdasarkan akurasi, presisi, recall, dan f-1 score dapat metode XGBoost menghasilkan nilai tertinggi untuk melakukan prediksi. Besar akurasi dengan menggunakan XGBoost adalah 90,89 %.

PENDAHULUAN

Pandemi COVID-19 yang terjadi pada tahun 2019 telah membuat beberapa perubahan dalam bidang pendidikan. Salah satunya adalah semakin banyaknya penggunaan teknologi pendidikan dalam pembelajaran akibat adanya pembatasan pembelajaran tatap muka dan meluasnya pembelajaran dalam jaringan (daring). Penggunaan teknologi tersebut dimulai dari pertemuan antara pendidik dan peserta didik yang menggunakan media online, pengumpulan tugas secara online, penyediaan forum diskusi secara online, hingga absensi siswa yang juga dihimpun secara online. Akhirnya data pendidikan yang selama ini tidak terekam dan masih menggunakan bukti fisik berubah menjadi lebih lengkap, sistematis, dan dapat diakses secara online. Dengan adanya pertumbuhan data pendidikan, kebutuhan untuk menemukan informasi yang produktif telah menjadi topik hangat di dunia penelitian (S.Kalaivani, 2017). Pertumbuhan data pendidikan tersebut harus dapat dimanfaatkan untuk mendukung peningkatan kualitas pembelajaran dan pendidikan.

Keberhasilan sistem pendidikan dapat dilihat dari kualitas pendidikan dari tahun ke tahun. Salah satu yang menjadi indikator kualitas pendidikan adalah tingkat kelulusan siswa. Perlu dilakukan evaluasi yang dilakukan oleh sekolah sebagai instansi pendidikan terkait dengan tingkat kelulusan siswa. Evaluasi dapat mendorong siswa untuk lebih giat belajar dan mendorong guru untuk lebih meningkatkan kualitas proses pembelajaran serta mendorong sekolah untuk lebih meningkatkan fasilitas dan kualitas

belajar siswa (Mahira, 2017). Sayangnya evaluasi proses belajar siswa masih belum terstruktur dan sangat bergantung pada penilaian subjektif guru. Hasil kelulusan siswa baru diketahui setelah siswa melakukan ujian dan pernyataan status 'lulus' atau 'tidak lulus' jarang dilakukan saat tengah pembelajaran. Akibatnya, tidak ada peringatan dini untuk siswa secara sistematis dan berdasarkan data terkait dengan evaluasi belajar siswa selama ini.

Faktor pertumbuhan data pendidikan dan perubahan dalam pendidikan kontemporer telah menyebabkan penggunaan berbagai teknik *data mining* untuk memantau kinerja siswa dengan berbagai metode investigasi untuk menganalisis dan menemukan informasi tersembunyi dalam sistem pendidikan (Ogor, 2007). *Data mining* adalah proses menemukan hubungan dalam data yang tidak diketahui oleh pengguna dan menyajikannya dengan cara yang dapat dipahami sehingga hubungan tersebut dapat menjadi dasar pengambilan keputusan (McLeod & Schell, 2007). *Data mining* pada data pendidikan biasa disebut dengan *educational data mining*. *Educational data mining* adalah salah satu metode yang dapat digunakan untuk mempelajari pola dan mengidentifikasi tren (Adekitam, 2019). Hasil yang diberikan oleh *data mining* dapat menjadi masukan untuk pembuat kebijakan pendidikan memberikan suatu keputusan berbasis data sehingga mendukung tujuan untuk meningkatkan efisiensi dan kualitas pengajaran dan pembelajaran. Penggunaan berbagai teknik *data mining* dapat dipandang sebagai perubahan sistematis dan positif yang signifikan jika dilihat dan berfungsi sebagai instrumen yang dapat membantu institusi pendidikan menunggu menemukan penyelesaian dari permasalahan selama ini (Barneveld, 2012).

Salah satu permasalahan yang dapat diselesaikan dengan *education data mining* adalah evaluasi, penilaian, dan pemantauan terhadap belajar siswa termasuk di dalamnya adalah prediksi kelulusan. Prediksi kelulusan yang dilakukan akan menghasilkan prediksi status 'lulus' atau 'tidak lulus' untuk setiap siswa pada setiap mata pelajaran. Hasil prediksi berdasarkan pada proses pembelajaran siswa pada paruh waktu pembelajaran. Data siswa diekstrak dan ditentukan model yang sesuai dari kebiasaan siswa, perilaku yang tidak sesuai, dan faktor-faktor lainnya yang berpengaruh terhadap hasil belajar. Selanjutnya dari model tersebut, jika dimasukkan nilai-nilai dari faktor yang dianalisis, *educational data mining* akan mengeluarkan status 'lulus' atau 'tidak lulus' untuk siswa tersebut. Prediksi kelulusan siswa diharapkan dapat menjadi tolak ukur dan motivasi untuk lebih meningkatkan belajar yang dilakukan siswa selama ini. Prediksi kelulusan setiap siswa diharapkan juga dapat menjadi peringatan dini untuk siswa yang diprediksi tidak lulus agar lebih mengantisipasi dan berhati-hati apabila proses belajar yang dilakukannya belum maksimal. Prediksi siswa yang tidak lulus pun dapat menjadi dasar untuk guru dan sekolah dalam memberikan stimulus positif agar siswa tidak putus asa. Pemberian stimulus di tengah masa akademik diharapkan dapat mengintervensi performa siswa sehingga dapat mengubah hasil prediksi.

Penelitian terkait dengan prediksi kelulusan telah dilakukan oleh beberapa peneliti sebelumnya. Penelitian yang dilakukan oleh Sumpena dkk memprediksi kelulusan siswa yang mengambil paket C. Metode yang digunakan pada penelitian Sumpena dkk adalah Naive Bayes yang menggunakan probabilitas dan statistik sebagai dasar untuk memprediksi kelulusan siswa. Prediksi kelulusan siswa mempunyai akurasi sebesar 97,20 %. Penelitian lain terkait prediksi kelulusan dilakukan oleh Sismadi dkk. Penelitian tersebut menggunakan algoritma Fuzzy C-Means untuk memprediksi tingkat kelulusan siswa. Sismadi dkk menggunakan *Sum of Square Error* untuk mengukur keefektifan algoritma dalam mengklusterisasi tingkat kelulusan siswa. Tingkat akurasi yang didapat adalah 78%. Penelitian yang dilakukan Bunkar dkk juga memprediksi performa siswa (Bunkar, 2012). Prediksi yang dilakukan menggunakan *decision tree* sebagai metode klasifikasi. Prediksi yang dilakukan dengan memprediksi siswa yang lulus dan gagal. Pada penelitian ini akan melakukan prediksi dengan membandingkan 3 metode *machine learning* yang berbeda, yaitu *XGBoost Classifier*, *Random Forest*, dan *Adaboost Classifier*. Hasil yang didapatkan dapat menjadi peringatan dini untuk siswa di tengah masa akademik sehingga siswa dapat meningkatkan performanya dan guru dapat memberikan stimulus untuk siswa berdasarkan data yang telah dianalisis. Akhirnya, presentasi kelulusan siswa dapat ditingkatkan dan performa siswa yang diprediksi tidak lulus dapat diberikan intervensi sehingga dapat mengurangi presentasi ketidaklulusan siswa.

METODE

Penelitian ini menerapkan model *data mining* yaitu *Knowledge Discovery in Database* (KDD). KDD adalah rangkaian proses untuk ekstraksi informasi potensial yang belum diketahui dan bersifat implisit dari sekumpulan data menjadi informasi yang mudah dipahami (Andayani, 2007). Pada KDD terdapat beberapa tahapan, yaitu *data selection*, *preprocessing*, *transformation*, *data mining*, *interpretation/evaluation* (Fayyad dkk, 1996). Tahapan KDD ditunjukkan pada Gambar 1.



GAMBAR 1. Tahapan *Knowledge Discovery in Database*

Data

Data yang digunakan pada penelitian ini berasal dari data dua sekolah menengah asal Portugal yang dapat diakses pada <https://archive.ics.uci.edu/> dengan nama “*Student Performance Data set*”. Data berisi kinerja siswa sekolah menengah pada dua mata pelajaran yaitu Bahasa Portugis dan matematika. Total sampel pada data berjumlah 1.044 sampel yang terdiri dari 395 sampel untuk mata pelajaran matematika dan 649 sampel untuk mata pelajaran Bahasa Portugis. Data diperoleh berdasarkan laporan sekolah dan angket. Penelitian ini menggunakan nilai akhir (dinotasikan dengan G_3) sebagai kriteria kelas yang akan diklasifikasi, yaitu ‘lulus’ untuk siswa yang mempunyai nilai $G_3 \geq 10$ dan ‘tidak lulus’ untuk siswa yang mempunyai nilai $G_3 < 10$. Berdasarkan kriteria tersebut persentase siswa dengan kategori lulus sebesar 67,09% atau sebanyak 265 siswa dan persentase siswa dengan kategori tidak lulus sebesar 32,91% atau sebanyak 130 siswa. Data terdiri dari 33 variabel. Rincian mengenai 33 variabel yang terdapat pada data penelitian ini ditunjukkan pada Tabel 1.

TABEL 1. Rincian Variabel Penelitian

Atribut	Deskripsi	Nilai
<i>School</i>	Nama Sekolah	GP: <i>Gabriel Pereira</i> MS : <i>Mousinho da Silveira</i>
<i>Sex</i>	Jenis Kelamin	'F': <i>Female</i> (Wanita) 'M': <i>Male</i> (Laki-laki)
<i>Age</i>	Umur	15-22 tahun
<i>Address</i>	Jenis Alamat Rumah	'U': <i>urban</i> (Perkotaan) 'R': <i>rural</i> (Pedesaan)
<i>Famsize</i>	Jumlah Anggota Keluarga	'LE3': kurang atau sama dengan 3 'GT3': lebih dari 3
<i>Pstatus</i>	Status Kehidupan Orangtua	'T' - <i>living together</i> (Tinggal Bersama) 'A' - <i>apart</i> (Terpisah)
<i>Medu</i>	Pendidikan Ibu	0: tidak ada, 1: pendidikan dasar (kelas 4), 2: kelas 5 – 9 3: pendidikan menengah 4: pendidikan tinggi
<i>Fedu</i>	Pendidikan Ayah	0: tidak ada, 1: pendidikan dasar (kelas 4), 2: kelas 5 – 9 3: pendidikan menengah 4: pendidikan tinggi
<i>Mjob</i>	Pekerjaan Ibu	<i>Teacher</i> (Guru) <i>Health</i> (Kesehatan)

Atribut	Deskripsi	Nilai
		<i>Services</i> (Pelayanan) <i>at_home</i> (Ibu rumah tangga) <i>Other</i> (lainnya)
Fjob	Pekerjaan Ayah	<i>Teacher</i> (Guru) <i>Health</i> (Kesehatan) <i>Services</i> (Pelayanan) <i>at_home</i> (Ibu rumah tangga) <i>Other</i> (lainnya)
Reason	Alasan untuk memilih sekolah	<i>Close to home</i> (dekat dengan rumah) <i>School reputation</i> (reputasi sekolah), <i>'course' preference</i> (preferensi) <i>'other'</i> (lainnya)
Guardian	Wali murid	<i>Mother</i> (Ibu) <i>Father</i> (ayah) <i>Other</i> (lainnya)
Traveltime	Waktu perjalanan pulang sekolah	1 : < 15 menit 2: 15 - 30 menit 3: 30 menit.- 1 jam 4: > 1 jam
Studytime	Waktu belajar mingguan	1: < 2 jam 2: 2 - 5 jam 3: 5 - 10 jam 4: > 10 jam
Failures	Banyaknya ketidakkululusan pada kelas sebelumnya	0, 1, 2, 3
Schollsup	Dukungan pendidikan ekstra	<i>Yes</i> (Ya) <i>No</i> (Tidak)
Famsup	Dukungan pendidikan keluarga	<i>Yes</i> (Ya) <i>No</i> (Tidak)
Paid	Kelas berbayar ekstra dalam mata pelajaran	<i>Yes</i> (Ya) <i>No</i> (Tidak)
Activities	Kegiatan ekstrakurikuler	<i>Yes</i> (Ya) <i>No</i> (Tidak)
Nursery	Bersekolah di Taman Kanak-Kanak	<i>Yes</i> (Ya) <i>No</i> (Tidak)
Higher	Keinginan melanjutkan ke pendidikan tinggi	<i>Yes</i> (Ya) <i>No</i> (Tidak)
Internet	Akses internet di rumah	<i>Yes</i> (Ya) <i>No</i> (Tidak)
Romantic	menjalani hubungan romantic	<i>Yes</i> (Ya) <i>No</i> (Tidak)
Famrel	kualitas hubungan keluarga	dari 1 - <i>very bad</i> (sangat buruk) sampai 5 - <i>excellent</i> (sangat baik)
Freetime	waktu luang sepulang sekolah	dari 1 - <i>very bad</i> (sangat buruk) sampai 5 - <i>excellent</i> (sangat baik)
Goout	pergi bersama teman-teman	dari 1 - <i>very bad</i> (sangat buruk) sampai 5 - <i>excellent</i> (sangat baik)
Dalc	konsumsi alkohol pada hari kerja	dari 1 - <i>very bad</i> (sangat buruk) sampai 5 - <i>excellent</i> (sangat baik)
Walc	konsumsi alkohol pada akhir pekan	dari 1 - <i>very low</i> (sangat rendah) sampai 5 - <i>very high</i> (sangat tinggi)
Health	status kesehatan saat ini	dari 1 - <i>very low</i> (sangat rendah) sampai 5 - <i>very high</i> (sangat tinggi)
Absences	jumlah absen di sekolah	0 - 93
G1	Nilai tahap pertama	0 – 20
G2	Nilai tahap kedua	0 – 20

Atribut	Deskripsi	Nilai
G3	Nilai akhir (<i>output target</i>)	0 - 20

Data Selection

Penelitian ini terbatas pada dataset mata pelajaran matematika dan tidak menggunakan dataset Bahasa Portugis dalam pembangunan modelnya sehingga sampel pada penelitian ini hanya berjumlah 395 siswa. Pengolahan data menggunakan Bahasa Python dalam *environment* Google Collab dan metode *machine learning* yang digunakan adalah metode klasifikasi.

Data Preprocessing dan Transformation

Sebelum menuju proses transformasi, data perlu dipersiapkan atau di-*preprocessing* terlebih dahulu untuk memaksimalkan hasil dari model *machine learning* yang terbaik (Suhanda dkk, 2020). *Preprocessing* adalah salah satu tahapan pada *data mining* yang salah satu tujuannya adalah untuk persiapan dan transformasi data ke dalam bentuk yang lebih sesuai (Alasadi dkk, 2017). Tujuan *preprocessing* data adalah untuk mereduksi ukuran data, menemukan hubungan antara data, normalisasi data, membuang pencilan dan mengekstrak fitur pada data. *Preprocessing* seringkali memberikan dampak yang signifikan pada kinerja *machine learning* khususnya untuk *supervised learning* termasuk di dalamnya untuk klasifikasi (Kotsiantis dkk, 2006).

Sebelum dilakukan normalisasi, data terlebih dahulu diubah nilainya karena jenis datanya yang berbeda-beda. Transformasi data mengubah data yang sebelumnya kategorik menjadi numerik. Selanjutnya, data yang digunakan perlu dinormalisasi agar semua variabel memiliki rentang nilai yang sama. metode yang akan digunakan untuk normalisasi adalah metode *min max scaler* yang akan menghasilkan rentang nilai antara 0 sampai 1. Formula dari *min max scaler* ditunjukkan pada (1)

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}, \quad (1)$$

dimana x_{scaled} adalah besaran nilai pada variabel x setelah dinormalisasi, x_{min} adalah nilai minimum variabel x , dan x_{max} adalah nilai maksimum variabel x (Raju dkk, 2020).

Data Mining

Data mining yang digunakan dalam penelitian ini berupa proses klasifikasi. Jaringan saraf tiruan (*neural network*) cenderung mengungguli performa semua algoritma *machine learning* lainnya baik untuk proses regresi maupun klasifikasi. Namun, dalam hal data terstruktur berjumlah kecil hingga menengah, algoritma berbasis *decision tree* dianggap yang terbaik di kelasnya saat ini. Metode *machine learning* untuk klasifikasi yang digunakan pada penelitian ini adalah tiga algoritma atau metode yang berbasis *decision tree* yaitu *Random Forest*, *XGBoost*, dan *Adaboost Classifier*. *Random Forest* adalah algoritma pembelajaran mesin yang dapat digunakan untuk berbagai tugas termasuk regresi dan klasifikasi. *Random Forest* termasuk metode *ensemble*, yang berarti bahwa model *Random Forest* terdiri dari sejumlah besar *decision tree* kecil (estimator) yang masing-masingnya dapat menghasilkan prediksi mereka sendiri-sendiri. Model *Random Forest* menggabungkan prediksi estimator tadi untuk menghasilkan prediksi yang lebih akurat. Pengklasifikasi *decision tree* standar memiliki kelemahan yaitu mereka rentan terhadap *overfitting* ke *training data*. Desain *ensemble Random Forest* memungkinkan *Random Forest* untuk mengkompensasi hal ini dan menggeneralisasi dengan baik untuk data diluar jangkauan *training data*, termasuk untuk data yang mengandung *missing value*. *Random Forest* juga bagus dalam menangani kumpulan data besar dengan dimensi tinggi dan tipe fitur heterogen (misalnya, jika satu kolom adalah kategorikal dan kolom lainnya numerik). *Random Forest* sangat baik untuk masalah klasifikasi tetapi sedikit kurang baik dalam masalah regresi. Berbeda dengan metode regresi linier, regresor *Random Forest* tidak dapat membuat prediksi di luar rentang data pelatihannya.

Metode kedua adalah *XGBoost* yang juga berbasis *decision tree* namun menggunakan kerangka kerja peningkatan gradien untuk optimasi dan iterasinya. *XGBoost* (*Extreme Gradient Boosting*) adalah penerapan lebih lanjut dari metode *Gradient Boosting Classifier*. *XGBoost* menggunakan prinsip

ensemble yaitu prinsip yang menggabungkan beberapa buah *tree* yang lemah menjadi sebuah model yang lebih kuat hingga akhirnya dapat menghasilkan prediksi yang baik. Metode terakhir adalah *AdaBoost*. *AdaBoost* yang juga dikenal dengan *Adaptive Boosting* adalah teknik dalam *machine learning* yang juga menggunakan prinsip *ensemble*. Algoritma yang paling umum digunakan dengan *AdaBoost* adalah *decision tree* dengan satu level yang berarti *decision tree* yang hanya terdiri atas 1 *split*. *Decision trees* ini juga disebut dengan *Decision Stumps*. Teknik *Boosting* bisa dikombinasikan dengan algoritma *classifier* lain untuk meningkatkan performa klasifikasi. Tentunya secara intuitif, penggabungan beberapa model akan berguna atau lebih baik jika model tersebut berbeda satu dengan yang lainnya. *Adaboost* adalah *meta-estimator* yang dimulai dengan memasang suatu *classifier* pada dataset asli dan kemudian memasang *classifier* berikutnya pada salinan dataset yang sama tetapi ditambahkan dengan proses pemberian atau penyesuaian bobot pada data-data yang salah diklasifikasi sehingga *classifier* tersebut lebih berfokus hanya pada data-data yang salah diklasifikasi sebelumnya dan memberikan hasil klasifikasi yang lebih baik pada *classifier* berikutnya. Tiga metode tersebut kemudian akan dibandingkan satu sama lain dengan melihat beberapa nilai atau metrik evaluasinya. Penerapan ketiga metode tersebut pada penelitian ini dengan menggunakan *environment* Google Colab yang berbahasa Python.

Interpretation/Evaluation

Setelah mendapat hasil klasifikasi dari tiga metode yang dijelaskan sebelumnya, selanjutnya dilakukan tahap evaluasi dengan menggunakan *Confusion Matrix*. *Confusion Matrix* adalah tabel yang berisikan banyaknya hasil klasifikasi yang diuji benar dan salah (Normawati, 2021). Bentuk umum *Confusion Matrix* ditunjukkan pada Tabel 2.

TABEL 2. Bentuk Umum *Confusion Matrix*

		Kelas Prediksi	
		Positif	Negatif
Kelas Sebenarnya	Positif	TP	FN
	Negatif	FP	TN

Keterangan:

- TP (*True Positive*) : banyaknya data uji kelas positif yang diprediksi sebagai kelas positif.
- TN (*True Negative*) : banyaknya data uji kelas negatif yang diprediksi sebagai kelas negatif.
- FP (*False Positive*) : banyaknya data uji kelas negatif yang diprediksi sebagai kelas positif.
- FN (*False Negative*) : banyaknya data uji kelas positif yang diprediksi sebagai kelas negatif.

Nilai dari *confussion matrix* yang digunakan untuk menghitung nilai akurasi, presisi, *recall*, dan *f1-score*. Formula untuk menghitung nilai akurasi, presisi, *recall*, dan *f1-score* ditunjukkan pada (2) – (5) secara berturut-turut (Fikri dkk, 2020).

$$Akurasi = \frac{TP + TN}{TP + FP + FN + TN} \tag{2}$$

$$Presisi = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$f1 - score = \frac{2 \times presisi \times recall}{presisi \times recall} \tag{5}$$

HASIL DAN PEMBAHASAN

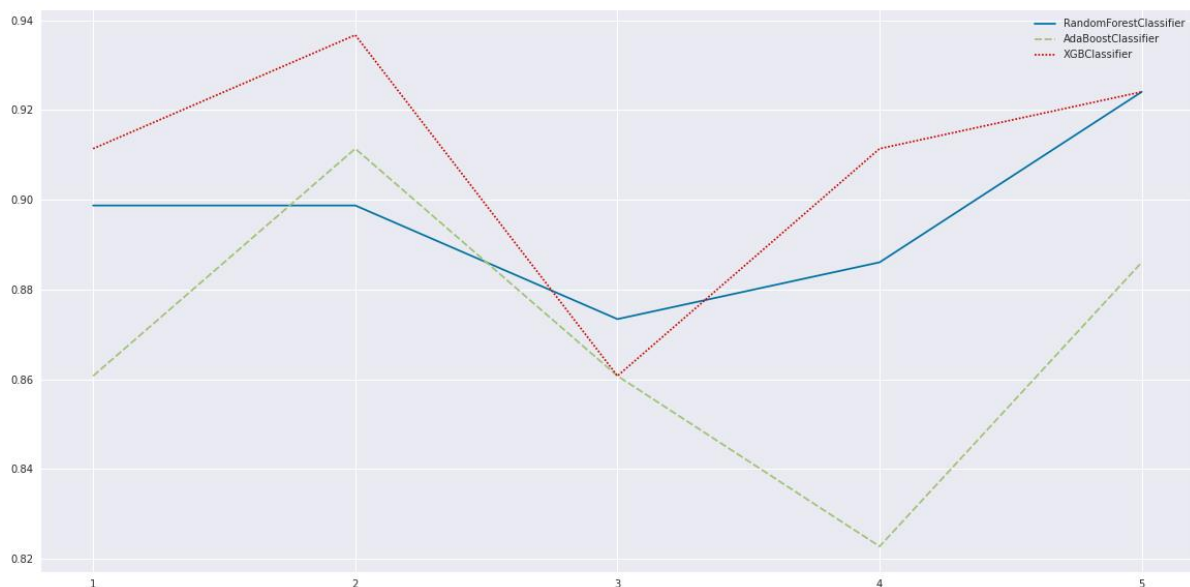
Educational Data Mining

Educational Data Mining adalah proses implementasi metode atau model *machine learning* beserta analisisnya terhadap data pendidikan. Setelah data Pendidikan pada penelitian ini diolah, ditransformasi dan kemudian dinormalisasi, data diimplementasikan ke dalam tiga model *machine learning* yaitu *Random Forest*, *AdaBoost*, dan *XGBoost*. Data yang diimplementasikan ke dalam tiga model *machine learning* tersebut dipartisi menjadi *training data* dan *test data* menggunakan *k-Fold Cross Validation* dengan menggunakan nilai $k = 5$. Setelah dilakukan penerapan tiga model pada data tersebut kemudian dilihat hasil nilai *accuracy* dari *training data*. Karena menggunakan *5-Fold Cross Validation*, diperoleh lima nilai akurasi yang berbeda-beda dari setiap jenis *fold*-nya seperti yang ditampilkan pada tabel 3.

TABEL 3. Nilai akurasi dari *training data* menggunakan *5-Fold Cross Validation*

	1	2	3	4	5
<i>XGBoost</i>	0.911392	0.936709	0.860759	0.911392	0.924051
<i>Random Forest</i>	0.886076	0.911392	0.886076	0.873418	0.936709
<i>Ada Boost</i>	0.860759	0.911392	0.860759	0.822785	0.886076

Agar dapat lebih terlihat jelas perbedaan nilai akurasi dari tiga metode yang digunakan pada penelitian ini, kelima nilai akurasi tersebut disajikan dalam grafik. Sumbu- x pada grafik menjelaskan tentang jenis *fold* yang terjadi pada *5-fold Cross Validation* (berarti ada lima kali perhitungan) dan sumbu- y berupa nilai akurasi (interval 0 sampai 1) dari *training data*. Grafik tersebut ditampilkan pada Gambar 2.



GAMBAR 2. Perbandingan akurasi dari *training data* menggunakan *5-Fold Cross Validation*

Tujuan dari dilakukannya hal ini adalah untuk melihat seberapa akurat setiap model atau metode yang telah diterapkan pada *training data*. Dari Gambar 2 dapat dilihat bahwa *XGBoost* mengungguli dua metode lainnya di empat perhitungan dari lima kali perhitungan yang telah dilakukan. Selanjutnya akan dilihat apakah *XGBoost* masih merupakan metode yang terbaik jika perhitungan diterapkan pada *testing data*. Namun kali ini pengukuran performa tiga model pada *testing data* menggunakan beberapa

metrik selain *accuracy* yaitu *precision*, *recall*, dan *f1-score*. Hasil evaluasi tersebut ditampilkan pada Tabel 4.

TABEL 4. Nilai *Accuracy*, *Precision*, *Recall*, dan *F1-Score* dari tiga model

	<i>test accuracy</i>	<i>test_precision</i>	<i>test_recall</i>	<i>test_f1-score</i>
<i>XGBoost</i>	0.908861	0.857673	0.869231	0.862443
<i>Random Forest</i>	0.896203	0.849850	0.838462	0.841974
<i>AdaBoost</i>	0.868354	0.826470	0.761538	0.789104

Dari hasil di atas, metode *XGBoost* merupakan metode yang semua nilai metriknya tertinggi dari ketiga model. Artinya pada penelitian ini diperoleh hasil bahwa metode yang akan dipilih adalah metode *XGBoost* karena metode ini secara mutlak mengungguli *Random Forest* dan *AdaBoost*.

Prediksi Kelulusan Siswa dengan Metode *XGBoost*

Hasil dari proses KDD pada penelitian ini adalah prediksi kelulusan siswa. Klasifikasi dibuat dengan membagi kelas dengan label ‘lulus’ untuk siswa yang mempunyai nilai $G3 \geq 10$ dan ‘tidak lulus’ untuk siswa yang mempunyai nilai $G3 < 10$. Hasil terbaik dari tiga metode yang diujicoba adalah dengan menggunakan *XGBoost*. Hasil tersebut dapat dipergunakan oleh guru untuk melihat apakah setiap siswa yang dididiknya akan diprediksi lulus atau tidak. Jika diprediksi tidak lulus, guru dapat memberikan intervensi kepada siswa tersebut baik dengan tambahan belajar, bimbingan, atau konsultasi kepada siswa. Guru dapat memberikan arahan positif kepada siswa agar tidak rendah diri dan kecewa terhadap hasil prediksi. Guru harus menjadi penyemangat dan pendorong untuk siswa agar hasil yang diprediksi dapat berubah menjadi lulus. Bagi siswa yang diprediksi lulus, guru dapat menjadi pendamping siswa agar tetap semangat dalam belajar dan tidak mengubah performa siswa agar hasil yang didapatkan sesuai dengan prediksi. Hal ini sesuai dengan penelitian yang dilakukan oleh Sinaga dan Handoko (2021) bahwa implementasi *data mining* dapat mempercepat pengolahan nilai kelulusan siswa dan membantu mendapatkan informasi yang berguna di masa yang akan datang.

PENUTUP

Kesimpulan

Salah satu pengaruh Pandemi COVID-19 adalah semakin terarsipnya data-data pendidikan akibat pembelajaran yang dilakukan secara daring. Data-data tersebut dapat digali untuk melakukan prediksi kelulusan siswa. Prediksi kelulusan siswa akan membantu guru dan penyelenggara pendidikan dalam membuat kebijakan sehingga pelayanan yang diberikan kepada siswa dapat meningkat. Prediksi kelulusan siswa menggunakan data pendidikan yang telah didapatkan sebelumnya sehingga *machine learning* dapat belajar. Pada penelitian ini menggunakan tiga metode *machine learning*, yaitu *Random Forest*, *AdaBoost*, dan *XGBoost*. Berdasarkan akurasi, presisi, *recall*, dan *f-1 score* dapat metode *XGBoost* menghasilkan nilai tertinggi untuk melakukan prediksi. Besar akurasi dengan menggunakan *XGBoost* adalah 90,89 %. Prediksi yang didapat diharapkan dapat menjadi panduan guru untuk melakukan intervensi dalam kegiatan belajar siswa dan menurunkan angka ketidakkelulusan pada siswa. Pengembangan penelitian selanjutnya dapat mengaplikasikan metode *machine learning* lainnya sehingga mendapat metode terbaik yang dapat meningkatkan hasil prediksi kelulusan siswa.

REFERENSI

- Adekitan, A. I., & Salau, O. (2019). The impact of engineering students' performance in the first three years on their graduation result using educational data mining. *Heliyon*, 5(2), e01250.
- Alasadi, S. A., & Bhaya, W. S. (2017). Review of data preprocessing techniques in data mining. *Journal of Engineering and Applied Sciences*, 12(16), 4102-4107.
- Andayani, S. (2007). Pembentukan cluster dalam Knowledge Discovery in Database dengan Algoritma K-Means. In *Seminar Nasional Matematika dan Pendidikan Matematika*.
- Bunkar, K., Singh, U. K., Pandya, B., & Bunkar, R. (2012, September). Data mining: Prediction for performance improvement of graduate students using classification. In *2012 Ninth International Conference on Wireless and Optical Communications Networks (WOCN)* (pp. 1-5). IEEE.
- Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, 17(3), 37-37.
- Fikri, M. I., Sabrila, T. S., & Azhar, Y. (2020). Perbandingan Metode Naïve Bayes dan Support Vector Machine pada Analisis Sentimen Twitter. *Smatika Jurnal*, 10(02), 71-76.
- Kotsiantis, S. B., Kanellopoulos, D., & Pintelas, P. E. (2006). Data preprocessing for supervised learning. *International journal of computer science*, 1(2), 111-117.
- Mahirah, B. (2017). Evaluasi belajar peserta didik (siswa). *Idarah: Jurnal Manajemen Pendidikan*, 1(2).
- McLeod, Jr.R. dan G.P. Schell.(2007). *Management Information System*. 10th ed. Pearson Education
- Normawati, D., & Prayogi, S. A. (2021). Implementasi Naïve Bayes Classifier Dan Confusion Matrix Pada Analisis Sentimen Berbasis Teks Pada Twitter. *J-SAKTI (Jurnal Sains Komputer dan Informatika)*, 5(2), 697-711.
- Ogor, E. N. (2007, September). Student academic performance monitoring and evaluation using data mining techniques. In *Electronics, robotics and automotive mechanics conference (CERMA 2007)* (pp. 354-359). IEEE.
- Raju, V. G., Lakshmi, K. P., Jain, V. M., Kalidindi, A., & Padma, V. (2020, August). Study the influence of normalization/transformation process on the accuracy of supervised classification. In *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)* (pp. 729-735). IEEE.
- S. Kalaivani, B. Priyadarshini, and B.S. Nalini (2017), Analyzing students' academic performance based on data mining approach. *International Journal of Innovative Research in Computer Science and Technology*, 194-197
- Sinaga, K., & Handoko, K. (2021). Implementasi Data Mining Untuk Memprediksi Kelulusan Siswa Dengan Metode Naïve Bayes. *Computer and Science Industrial Engineering (COMASIE)*, 4(6), 97-107.
- Suhanda, Y., Kurniati, I., & Norma, S. (2020). Penerapan Metode Crisp-DM Dengan Algoritma K-Means Clustering Untuk Segmentasi Mahasiswa Berdasarkan Kualitas Akademik. *Jurnal Teknologi Informatika dan Komputer*, 6(2), 15.
- Sumpena, J., & Kurnia, N. (2019). Analisis Prediksi Kelulusan Siswa PKBM Paket C Dengan Metoda Algoritma Naïve Beyes. *Jurnal TEDC*, 13(2), 127-133.
- Sismadi, S., & Kusnadi, Y. (2018). PREDIKSI TINGKAT KELULUSAN SISWA ELEARNING BERBASIS ALGORITMA FUZZY C-MEANS. *Techno Nusa Mandiri: Journal of Computing and Information Technology*, 15(1), 1-6.
- Van Barneveld, A., Arnold, K. E., & Campbell, J. P. (2012). Analytics in higher education: Establishing a common language. *EDUCAUSE learning initiative*, 1(1), 1-11.