BRAHMANA: Jurnal Penerapan Kecerdasan Buatan
Terakreditasi Nomor 204/E/KPT/2022 | Vol. 4, No. 2, Juni (2023), pp. 173-183

Prediksi Kelulusan Siswa Sekolah Menengah Pertama Menggunakan Machine Learning

Agusti Frananda Alfonsus Naibaho[1], Amalia Zahra[2]
[1,2]Computer Science Department, BINUS Graduate Program, Master of Computer Science, Universitas Bina Nusantara, Jakarta, Indonesia
E-mail: [1]agusti.anaibaho@binus.ac.id, [2]amalia.zahra@binus.edu

Abstract

In recent years, there have been students who have not graduated on time at Lubuk Alung 1 Public Junior High School. This statement is supported by graduation data from SMP Negeri 1 Lubuk Alung. Therefore it is necessary to predict student graduation status to identify factors that influence student graduation, which can also be used to help schools solve problems more easily. To overcome this problem, researchers predict student graduation based on student graduation information. The attributes used are personal data related to students, student academic data, and data related to the work of students' parents. Researchers obtained data on student graduation from schools that had been recapitulated. The classification algorithms used are decision tree, random forest, and extreme gradient boosting with grid searchCV and k-fold=5. Predictive accuracy using the random forest algorithm outperforms other methods with a value of 99.5%.

Keywords: *Graduation Prediction, Students, Machine Learning, Data Mining*

Abstract

Dalam beberapa tahun terakhir, terdapat siswa yang lulus tidak tepat waktu di Sekolah Menengah Pertama Negeri 1 Lubuk Alung. Pernyataan ini didukung oleh data kelulusan dari Sekolah Menengah Pertama Negeri 1 Lubuk Alung. Oleh karena itu perlu dilakukan prediksi status kelulusan siswa untuk mengidentifikasi faktor yang mempengaruhi kelulusan siswa, yang juga dapat digunakan untuk membantu sekolah memecahkan masalah menjadi lebih mudah. Untuk mengatasi masalah tersebut, peneliti memprediksi kelulusan siswa berdasarkan informasi kelulusan siswa. Atribut yang digunakan adalah data pribadi yang berhubungan dengan siswa, data akademik siswa, dan data yang berhubungan dengan pekerjaan orang tua siswa. Peneliti memperoleh data kelulusan siswa dari sekolah yang telah direkapitulasi. Algoritma klasifikasi yang digunakan adalah decision tree, random forest, dan extreme gradient boosting dengan grid searchCV dan k-fold=5. Akurasi prediksi menggunakan algoritma random forest mengungguli metode lainnya dengan nilai 99,5%.

Keywords: *Graduation Prediction, Students, Machine Learning, Data Mining*

1. Pendahuluan

Data siswa merupakan salah satu informasi yang paling penting di bidang akademik. Setiap institusi pendidikan menyimpan data siswa pada *database*. Data siswa memiliki beberapa informasi yang berguna dimana biasanya tidak hanya tercantum transkrip nilai siswa, profil siswa, dan beberapa data lainnya, namun juga terdapat pola yang dapat digunakan sebagai bahan analisis. Kumpulan data siswa dapat digunakan untuk memprediksi lama waktu siswa dalam menyelesaikan pendidikan serta informasi mengenai performa siswa [1]. Memprediksi performa siswa berguna untuk sekolah yang dapat digunakan sebagai bahan evaluasi meningkatkan pembelajaran dan proses pengajaran. Terutama memprediksi lama waktu belajar siswa adalah hal yang sangat penting untuk sekolah dalam membantu siswa mengatur rencana studi serta memberikan pelajaran tambahan.

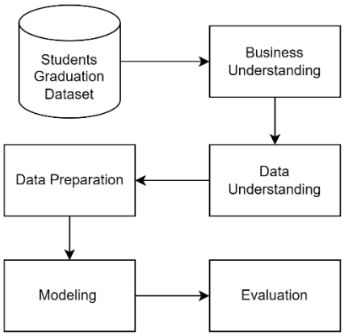
ISSN: 2715-9906

Lama waktu belajar menjadi salah satu indikator yang penting dalam sistem pendidikan di Indonesia. Lama waktu belajar siswa digunakan oleh institusi pendidikan untuk memberikan penilaian terhadap sekolah mengenai gambaran kinerja sekolah sebagai alat pembinaan, pengembangan, dan peningkatan mutu serta menentukan tingkat kelayakan sekolah sebagai lembaga penyelenggara pelayanan pendidikan. Untuk dapat melakukan prediksi mengenai performa serta durasi siswa dalam menyelesaikan pendidikan, maka pada data siswa dapat diterapkan *machine learning*. *Machine learning* adalah cabang ilmu yang mencakup perancangan dan pengembangan algoritma yang memungkinkan komputer untuk mengembangan perilaku berdasarkan data yang diberikan. Model algoritma yang sesuai dihasilkan oleh pengalaman (*experience*) dan proses pembuatan model algoritma merupakan proses pembelajaran (*learning*) otomatis oleh mesin [2].

Telah dilakukan beberapa penelitian menggunakan *machine learning* yang diterapkan dalam dunia pendidikan. Algortima *decision tree* diterapkan pada penelitian tentang memprediksi kinerja akademik [3], menilai atribut pembelajaran jarak jauh yang mempengaruhi performa mahasiswa [4], dan memprediksi kepuasan mahasiswa terhadap layanan akademik universitas [5]. Hasil yang diperoleh menunjukkan hasil evaluasi yang baik dimana *decision tree* mampu menghasilkan nilai akurasi yang sangat baik. Dalam penelitian lain, algoritma *random forest* diterapkan memprediksi performa akademik siswa sekolah menengah pertama dan siswa sekolah menengah atas [6], memprediksi kemungkinan penerimaan mahasiswa pasca sarjana [7][8], mengidentifikasi siswa yang memiliki kemungkinan putus sekolah [9], dan memprediksi rata-rata nilai kelulusan *CGPA* untuk mendeteksi performa akademik yang buruk [10]. Hasil yang diperoleh dari penelitian menggunakan *random forest* menunjukkan hasil evaluasi yang baik dan mampu mengidentifikasi variabel yang paling berkorelasi dengan tugas prediksi. Algoritma *extreme gradient boosting* diaplikasikan pada penelitian tentang prediksi performa mahasiswa [11], prediksi prestasi mahasiswa [12], dan prediksi status penyelesaian mahasiswa yang telah mencapai lama studi maksimal [13]. Hasil yang diperoleh menunjukkan hasil evaluasi yang baik. Berdasarkan beberapa penelitian yang dilakukan, algoritma *decision tree*, *random forest*, dan *extreme gradient boosting* dinilai akan cukup mampu menangani kasus prediksi kelulusan siswa.

2. Metodologi Penelitian

Bahan penelitian yang digunakan adalah *dataset* kelulusan siswa Sekolah Menengah Pertama Negeri 1 Lubuk Alung yang telah direkapitulasi dalam tiga tahun terakhir yang diperoleh dari pihak sekolah. Tahapan penelitian dimulai dengan memahami nilai bisnis dari penelitian, pemahaman data, persiapan data, pemodelan, dan evaluasi.



Gambar 1. Tahapan Penelitian

2.1. *Machine Learning*

Machine learning adalah cabang ilmu *artificial intelligence* yang menggunakan berbagai statistik, teknik probabilitas, dan optimasi yang memungkinkan komputer untuk belajar dari contoh masa lalu dan mendeteksi pola yang sulit dibedakan dari sejumlah besar data atau kompleks [9]. Kemampuan ini sangat cocok untuk berbagai

pengaplikasian yang bergantung pada pengukuran yang kompleks. Secara luas proses *machine learning* terdiri dari enam proses utama [14] yaitu:

1. *Gathering data*, merupakan proses dasar dari *machine learning*, namun proses ini sangat penting karena kualitas dan kuantittas data dari proses ini akan membantu menentukan kualitas model prediksi.
2. *Data preparation*, merupakan proses mempersiapkan data sehingga dapat digunakan untuk proses *training machine learning*.
3. *Choosing a model*, merupakan proses pemilihan model yang relevan dengan studi penelitian.
4. *Training*, merupakan proses yang menggunakan data dalam perkembangan untuk meningkatkan kemampuan model untuk memprediksi.
5. *Evaluation* dan *tuning parameter*, merupakan proses untuk memeriksa keakuratan model dalam melakukan prediksi. Setelah proses evaluasi biasanya dilakukan *tuning parameter* untuk melihat apakah berdasarkan parameter yang dimodifikasi akan menyebabkan peningkatan atau penurunan pada hasil evaluasi.
6. *Prediction*, merupakan proses menggunakan data untuk memberikan jawaban terhadap pertanyaan.

2.2. *Data Mining*

Data mining adalah proses menemukan pola yang berpotensi dan berguna menggunakan kumpulan data yang besar. Proses *data mining* menggunakan ilmu matematika, *machine learning*, statistik, dan *artificial intelligence* untuk mengekstrak informasi tentang kemungkinan yang terjadi di masa mendatang [15]. *Data mining* merupakan teknik yang digunakan untuk menemukan pengetahuan yang tersembunyi dan hubungan yang tidak terduga antara data. *Data mining* disebut juga *Knowledge Discovery in Database* (*KDD*) dimana merupakan suatu teknik untuk menggali informasi berharga yang tidak diketahui sebelumnya pada kumpulan data yang sangat besar [16].

2.3. *Classification*

Classification merupakan salah satu *task* yang dianggap sebagai pendekatan *supervised learning* dalam *machine learning* dimana pembelajaran dengan program komputer dilakukan pada *input* data yang diberikan. *Classification* adalah proses menemukan model atau fungsi yang membedakan kelas atau konsep data [17]. Model diturunkan berdasarkan analisa sekumpulan *data train* yang label kelasnya diketahui. Berdasarkan pembelajaran ini, model akan mengklasifikasian hasil yang belum diperoleh sebelumnya. Proses klasifikasi didasarkan pada komponen kelas (*class*) prediktor (*predictor*), *dataset* pelatihan (*training dataset*), dan *dataset* pengujian (*testing dataset*) [17].

2.4. *Decision Tree*

Decision tree adalah sebuah metode klasifikasi yang berbentuk seperti pohon yang memiliki aturan-aturan. Atribut yang dipilih pada *decision tree* menghasilkan partisi dengan data yang lebih seragam dan dapat menghasilkan pohon keputusan yang sederhana dengan perulangan yang sedikit. Sebuah *decision tree* terdiri dari sekumpulan aturan yang bertujuan untuk membagi sejumlah populasi yang heterogen menjadi lebih kecil dan lebih homogen dengan memperhatikan variabel tujuan [18]. Cara kerja klasifikasi yang dilakukan menggunakan *decision tree* terdiri dari *internal node* yang menyatakan pengujian terhadap suatu atribut, setiap cabang menyatakan *output* dari pengujian tersebut, dan *leaf node* menyatakan kelas-kelas atau pembagian kelas. *Node* teratas pada *decision tree* disebut sebagai *root node* dimana biasanya memiliki pengaruh terbesar pada suatu kelas tertentu. *Decision tree* melakukan strategi pencarian secara *top down*. Untuk proses mengklasifikasikan data yang tidak diketahui, nilai atribut akan diuji dengan cara melacak jalur dari *root node* sampai *leaf node* dan kemudian akan diprediksi kelas yang dimilih oleh suatu data baru tertentu.

2.5. *Random Forest*

Random forest adalah sebuah metode *ensemble* untuk meningkatkan akurasi metode klasifikasi dengan cara mengkombinasikan metode klasifikasi. *Random forest* merupakan suatu metode klasifikasi yang berisi koleksi dari pohon klasifikasi (*decision tree*), dimana pada setiap *decision tree* telah dilakukan *training* menggunakan *sample* individu dan setiap atribut dipecah pada *tree* yang dipilih antara atribut *subset* yang bersifat acak (*random*). *Random forest* dikembangkan dari metode *CART* yaitu dengan menerapkan metode *bootstrap aggregating* (*bagging*) dan *random feature selection* [19]. *Random forest* mempunyai seleksi fitur dimana mampu mengambil fitur terbaik sehingga dapat meningkatkan performa terhadap model klasifikasi. *Random forest* dikembangkan dengan ide bahwa perlu terdapat penambahan *layer* pada proses *resampling* acak pada *bagging*. Selain itu data *sample* yang diambil secara acak untuk membentuk *decision tree*, variabel prediktor juga diambil secara acak dan baru dipilih sebagai pemilah terbaik saat penentuan pemilah pohon [20].

2.6. *Extreme Gradient Boosting*

Extreme gradient boosting atau *xgboost* merupakan sebuah metode pengembangan dari *gradient boosting*. *Gradient boosting* adalah algoritma yang dapat menemukan solusi optimal pada masalah regresi, klasifikasi, dan *ranking*. Konsep dasar dari *xgboost* adalah menyesuaikan parameter pembelajaran secara berulang untuk menurunkan *loss function* [21]. *Xgboost* bekerja berdasarkan *gradient boosting decision tree* dimana terdapat kumpulan *decision tree* yang pembangunan pohon berikutnya bergantung pada pohon sebelumnya [21]. *Xgboost* merupakan metode yang memprediksi *error* dari model sebelumnya. Pohon pertama pada *xgboost* cenderung lemah dalam melakukan klasifikasi. Penambahan pohon dilakukan dengan tujuan tidak terdapat lagi perbaikan *error* yang dapat dilakukan. Hasil prediksi akhir *xgboost* biasanya berupa penjumlahan hasil prediksi dari setiap pohon regresi [22]. Dalam *xgboost* diperlukan fungsi objektif yang berguna untuk menilai seberapa bagus model didapatkan sesuai dengan *data train*. Karakteristik utama dalam fungsi *xgboost* adalah terdapat dua fungsi objektif yang terdiri dari dua bagian yaitu nilai pelatihan yang hilang (*loss function*) dan nilai regularisasi [22].

3. Hasil dan Pembahasan

3.1. *Business Understanding*

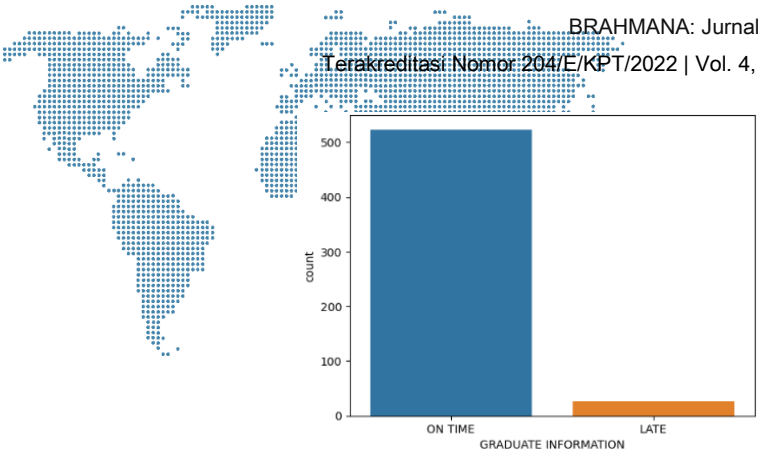
Tahap ini bertujuan untuk memahami nilai bisnis yang akan diperoleh berdasarkan penelitian yang dilakukan. Berdasarkan pemahaman dan analisis yang telah dilakukan, permasalahan yang dihadapi oleh Sekolah Menengah Pertama Negeri 1 Lubuk Alung adalah setiap tahunnya selalu terdapat siswa yang lulus tidak tepat waktu. Oleh karena itu pihak sekolah bermaksud untuk melakukan prediksi kelulusan siswa guna mengetahui penyebab keterlambatan kelulusan siswa, terutama variabel-variabel yang paling mempengaruhi kelulusan siswa berdasarkan data yang tersedia.

3.2. *Data Understanding*

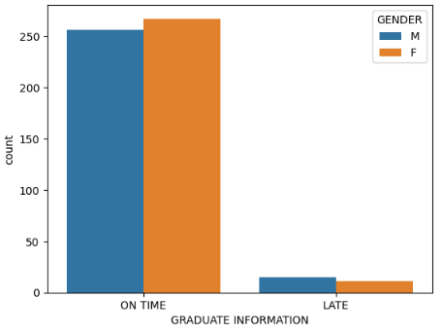
Tahapan yang dilakukan setelah semua data terkumpul dan telah dilakukan pemahaman nilai bisnis adalah pemahaman data (*data understanding*). Pada tahap ini, pemahaman data dilakukan dengan mengeksplorasi dan memverifikasi kualitas data yang digunakan. Hasil akhir dari tahap ini adalah mampu memahami data dan menemukan wawasan awal terhadap data yang digunakan.

3.2.1. Eksplorasi Data Kelulusan Siswa

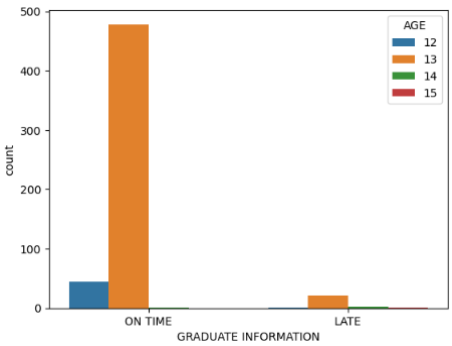
Pada tahap eksplorasi data, dilakukan eksplorasi dengan menggunakan visualisasi berdasarkan variabel yang terdapat pada data kelulusan siswa. Visualisasi ini digunakan sebagai bahan untuk memahami kondisi *dataset* yang digunakan. Beberapa hasil dari eksplorasi data kelulusan siswa dapat dilihat pada Gambar 2, 3, 4, 5, 6, 7, dan 8.



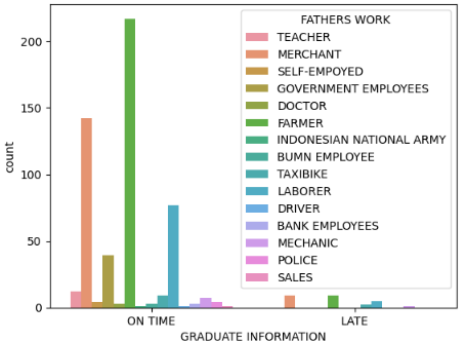
Gambar 2. Jumlah Siswa Berdasarkan Informasi Kelulusan Siswa



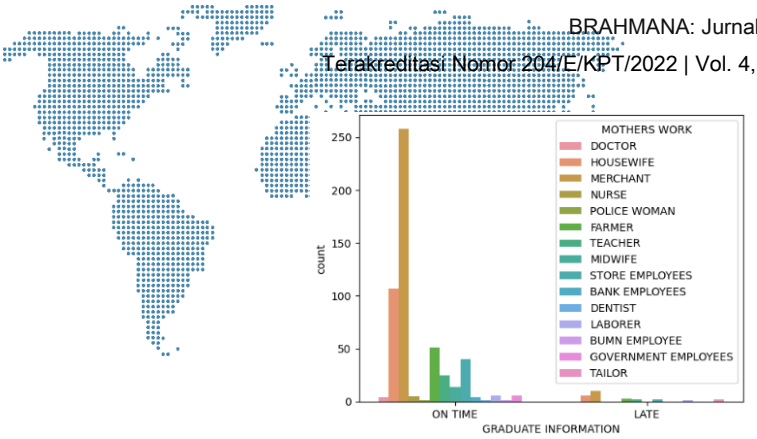
Gambar 3. Jumlah Siswa Berdasarkan Jenis Kelamin Siswa



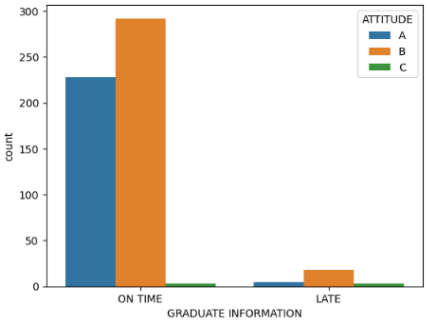
Gambar 4. Jumlah Siswa Berdasarkan Usia Siswa



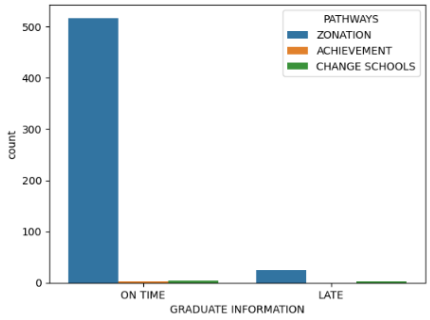
Gambar 5. Jumlah Siswa Berdasarkan Pekerjaan Ayah Siswa



Gambar 6. Jumlah Siswa Berdasarkan Pekerjaan Ibu Siswa



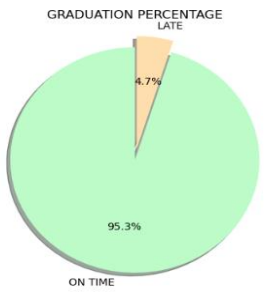
Gambar 7. Jumlah Siswa Berdasarkan Nilai Sikap Siswa



Gambar 8. Jumlah Siswa Berdasarkan Jalur Masuk Siswa

3.2.2. Verifikasi Kualitas Data

Tugas pertama dari tahap ini adalah memeriksa apakah terdapat data yang hilang (*missing value*) menggunakan fungsi *isna* dan *isnull*. Berdasarkan hasil pemeriksaan *missing value*, ditemukan bahwa tidak terdapat data yang hilang. Lalu berdasarkan eksplorasi data pada Gambar 2, ditemukan bahwa terdapat daya yang tidak seimbang. Dilakukan pemeriksaan keseimbangan data menggunakan visualisasi yang menghasilkan persentase kelulusan siswa berdasarkan informasi kelulusan.



Gambar 9. Persentase Jumlah Kelulusan Siswa

Berdasarkan visualisasi yang dilakukan, terlihat bahwa data yang digunakan tidak seimbang (*imbalance data*). Penggunaan data yang tidak seimbang dapat mempengaruhi algoritma yang digunakan. Apabila algoritma memprediksi bahwa semua siswa lulus tepat waktu (*on time*), maka akurasi yang diperoleh adalah 95,3% dan hasil akan menunjukkan bahwa algoritma yang digunakan tidak mampu memprediksi dengan tepat jika data yang digunakan tidak seimbang.

3.3. *Data Preparation*

Pada tahap ini dilakukan dengan mengatasi ketidakseimbangan data (*imbalance data*) dan menghapus variabel yang dianggap tidak relevan berdasarkan pemahaman data.

3.3.1. Mengatasi Ketidakseimbangan Data

Berdasarkan hasil proses pemahaman data, ditemukan bahwa data yang digunakan dalam penelitian ini tidak seimbang (*imbalance*). Pada tahap ini dilakukan proses mengatasi ketidakseimbangan data menggunakan metode *random under sampling* dan metode *random over sampling*. Hasil dari proses mengatasi ketidakseimbangan data dapat dilihat pada Tabel 1 berikut.

Tabel 1. Hasil Mengatasi Ketidakseimbangan Data

No	Kondisi Data	*On Time*	*Late*
1	Original Data	523	26
2	*Random Under Sampling*	26	26
3	*Random Over Sampling*	523	523

Mengacu pada hasil proses *random sampling* pada Tabel 1, terdapat data yang saat ini seimbang. Pada metode *random under sampling*, beberapa data secara acak dihapus dari kelas mayoritas (kelas '*on time'*) sehingga dihasilkan pengurangan data pada kelas tepat waktu (*on time*) dan pada metode *random over sampling*, beberapa data ditambahkan dari kelas minoritas (kelas '*late'*) sehingga terdapat penambahan data pada kelas terlambat (*late*). Namun pada *random under sampling*, data yang digunakan berkurang secara signifikan sehingga dikhawatirkan dapat menghilangkan pengklasifikasi (*classifier*) yang sangat penting. Untuk alasan ini, data dari proses *random over sampling* akan digunakan untuk proses selanjutnya dalam penelitian ini.

3.3.2. Eliminasi Variabel yang Tidak Relevan

Berdasarkan hasil proses pemahaman data, ditemukan bahwa variabel nomor (atribut '*No'*) dalam data tidak relevan dengan penelitian ini sehingga akan dihapus dari data. Kemudian diperlukan ekstraksi data dengan mengubah data menjadi bentuk yang lebih mudah dipahami dengan proses penghilangan data yang tidak diperlukan. Proses pelabelan data nominal menjadi numerik dilakukan untuk membedakan kategori berdasarkan variasi data. Selain itu seleksi fitur dilakukan dengan menggunakan metode korelasi *Pearson* dengan *matrix correlation* untuk menghasilkan korelasi antar variabel. Berdasarkan hasil yang diperoleh ditemukan bahwa variabel jenis kelamin (*gender*) tidak terlalu berpengaruh, sehingga variabel jenis kelamin akan dihapus. Atribut yang akan digunakan untuk prediksi klasifikasi ditunjukkan pada Tabel 2 berikut.

Tabel 2. Atribut Data yang Digunakan untuk Prediksi Klasifikasi

No	Atribut	Tipe Atribut
1	*Islamic Religious Education*	*Numerical/Continuous*
2	*Civic Education*	*Numerical/Continuous*
3	*Indonesian*	*Numerical/Continuous*
4	*English*	*Numerical/Continuous*
5	*Mathematics*	*Numerical/Continuous*

No	Atribut	Tipe Atribut
6	*Natural Sciences*	*Numerical/Continuous*
7	*Social Sciences*	*Numerical/Continuous*
8	*Art and Culture*	*Numerical/Continuous*
9	*Physical Education*	*Numerical/Continuous*
10	*Enterpreneurship*	*Numerical/Continuous*
11	*Attitude*	*Nominal/Discrete*
12	*Pathways*	*Nominal/Discrete*
13	*Age*	*Numerical/Continuous*
14	*Number of Siblings*	*Numerical/Continuous*
15	*Position of Siblings*	*Numerical/Continuous*
16	*Father's Work*	*Nominal/Discrete*
17	*Father's Income* (*IDR*)	*Numerical/Continuous*
18	*Mother's Work*	*Nominal/Discrete*
19	*Mother's Income* (*IDR*)	*Numerical/Continuous*
20	*Graduate Information*	*Nominal/Discrete*

3.4. *Modeling* dan *Hyperparameter Tuning*

Algoritma *decision tree*, *random forest*, dan *extreme gradient boosting* akan digunakan dalam proses model prediksi klasifikasi. Mengacu pada penelitian sebelumnya, algoritma *decision tree*, *random forest*, dan *extreme gradient boosting* menunjukkan hasil yang baik dalam tugas prediksi klasifikasi. Pada penelitian ini, sebelumnya akan dilakukan *hyperparameter tuning* dengan menggunakan *grid searchCV* dengan *k-fold=5* untuk mendapatkan parameter terbaik yang akan diterapkan pada algoritma yang digunakan. Hasil dari proses *hyperparameter tuning* algoritma *machine learning* yang digunakan dapat dilihat pada Tabel 3, 4, dan 5 berikut.

Tabel 3. Hasil *Hyperparameter Tuning* dari *Decision Tree*

Parameter	*Grid SearchCV Values*	*Best Parameter*
max_depth	4, 5, 6, 7, 8	8
criterion	*gini*, *entropy*	*entropy*
min_samples_leaf	2, 3, 4, 5, 6, 7, 8, 9	4
min_samples_split	2, 3, 4, 5, 6, 7, 8	8

Tabel 4. Hasil *Hyperparameter Tuning* dari *Random Forest*

Parameter	*Grid SearchCV Values*	*Best Parameter*
n_estimators	100, 200, 300	100
max_depth	*None*, 1, 2, 3, 4, 5, 6, 7, 8	*None*
criterion	*gini*, *entropy*	*entropy*
min_samples_leaf	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	2
max_features	*auto*, *sqrt*, log^2	*auto*

Tabel 5. Hasil *Hyperparameter Tuning* dari *Extreme Gradient Boosting*

Parameter	*Grid SearchCV Values*	*Best Parameter*
n_estimators	100, 200, 300	100
max_depth	4, 5, 6, 7, 8	6
min_child_weight	1, 2, 3, 4, 5, 6, 7	1
eta (*learning_rate*)	0.025, 0.05, 0.1, 0.2, 0.3	0.2
gamma	0, 0.1, 0.2, 0.3, 0.4, 1.0, 1.5, 2.0	1.0
subsample	0.15, 0.5, 0.75, 1.0	1.0
colsamples_bylevel	log^2, *sqrt*, 0.25, 1.0	0.25

Hasil dari proses *hyperparameter tuning* didapatkan dari *grid searchCV* dengan melakukan pencarian menyeluruh terhadap parameter yang diuji. Proses *hyperparameter tuning* menggunakan validasi *k-fold=5* yang digunakan untuk mengevaluasi kinerja model sebanyak 5 iterasi pada proses *grid searchCV* setiap parameter. Hasil proses *hyperparameter tuning* selanjutnya digunakan untuk menentikan prediksi klasifikasi.

3.5. Evaluasi Prediksi Klasifikasi

Pada tahap prediksi klasifikasi, siswa yang lulus tepat waktu (*on time*) dan yang lulus terlambat (*late*) akan diklasifikasikan menggunakan algoritma yang diajukan. Model evaluasi yang digunakan adalah dengan menghasilkan nilai akurasi (*accuracy*) dan nilai *ROC-AUC* dari tiap algoritma. Hasil evaluasi dari prediksi klasifikasi menggunakan ketiga algoritma ditunjukkan pada Tabel 6 berikut.

Tabel 6. Hasil Evaluasi Prediksi Klasifikasi Kelulusan Siswa

Algoritma	*Accuracy*	*ROC-AUC*
Decision Tree	95,238%	95,098%
Random Forest	99,523%	99,509%
Extreme Gradient Boosting	99,047%	99,019%

Hasil evaluasi prediksi klasifikasi kelulusan siswa dengan menggunakan algoritma *random forest* menunjukkan nilai akurasi sebesar 99,5%, sedangkan hasil evaluasi yang diperoleh menggunakan algoritma *extreme gradient boosting* adalah sebesar 99,0%, sedikit lebih rendah yaitu 0,5% dibandingkan akurasi algoritma *random forest*. Hasil akurasi prediksi klasifikasi terendah berada pada algoritma *decision tree* yang memiliki akurasi 95,2%. Mengacu pada hasil yang akurat, dapat disimpulkan bahwa variabel yang digunakan dapat memberikan hasil yang meyakinkan. Selain itu nilai *ROC-AUC* diatas 90% menunjukkan bahwa setiap *classifier* dapat dengan tepat membedakan kelas mayoritas siswa yang lulus terlambat (*late*) dan siswa yang lulus tepat waktu (*on time*). Selain itu juga digunakan *feature importance* yang menunjukkan hubungan variabel yang digunakan dalam mempengaruhi hasil prediksi kelulusan siswa.

Gambar 10. Hasil *Feature Importance*

Berdasarkan Gambar 10, nilai kewirausahaan (*enterpreneurship*) merupakan variabel yang sangat mempengaruhi kelulusan siswa, lalu diikuti oleh variabel matematika (*mathematics*), bahasa Indonesia (*Indonesian*), dan ilmu pengetahuan alam (*natural sciences*) serta beberapa variabel lainnya. Oleh karena itu pihak sekolah perlu melakukan evaluasi berupa perbaikan cara memberikan pengajaran pada mata pelajaran seperti kewirausahaan, matematika, bahasa Indonesia, dan ilmu pengetahuan alam.

4. Kesimpulan

Berdasarkan uji prediksi klasifikasi kelulusan pada siswa Sekolah Menengah Pertama Negeri 1 Lubuk Alung dengan menghasilkan nilai akurasi sebesar 99,5%, maka atribut data pribadi yang berhubungan dengan siswa, data akademik siswa, dan data yang berhubungan dengan pekerjaan orang tua siswa terbukti cukup efektif dalam memprediksi kelulusan siswa menggunakan *machine learning*. Selain itu penggunaan *hyperparameter tuning* dapat mempermudah dalam menghasilkan parameter terbaik berdasarkan data kelulusan siswa yang digunakan. Terlepas dari hasil tersebut, terdapat beberapa kesalahan dalam proses prediksi klasifikasi kelulusan siswa yang disebabkan oleh cara mengatasi ketidakseimbangan data karena data yang digunakan sebagian besar merupakan duplikasi dari data siswa yang lulus terlambat. Hasil prediksi menunjukkan kecenderungan yang tinggi bahkan mendekati sempurna, sehingga menimbulkan ambiguitas dalam proses penyeimbangan data. Akibatnya dapat dikatakan bahwa terdapat data yang tidak unik di sebagian besar data terbaru yang menyebabkan klasifikasi berulang. Untuk mengatasi kesalahan dalam penyeimbangan data, diperlukan penelitian lebih lanjut untuk menganalisis penyeimbangan data agar data yang digunakan tidak memiliki kecenderungan berulang dan memberikan hasil yang baik. Selain itu beberapa aspek lainnya yang berkaitan dengan siswa juga dapat diteliti pada penelitian selanjutnya seperti gaya hidup siswa dan jumlah absensi siswa.

Daftar Pustaka

[1] T. Handayani, L. Hiryanto, "Predicting and Analyzing the Length of Study-Time Using Support Vector Machine (Case Study: Computer Science Students)", *ComTech Computer, Mathematics, and Engineering Applications*, Vol. 8, No. 2, 2017, pp. 107-114.

[2] B. Wu, C. Zheng, "An Analysis of the Effectiveness of Machine Learning Theory in the Evolution of Education and Teaching", *Hindawi Journal*, Northeast Normal University (China), 2021, pp. 1-10.

[3] I. O. Muraina, E. A. Aiyegbusi, S. O. Abam, "Decision Tree Algorithm Use in Predicting Students' Academic Performance in Advanced Programming Course", *International Journal of Higher Education Pedagogies*, Vol. 3, No. 4, 2022, pp. 13-23.

[4] H. Mohammad, Abu-Dalbouh, "Application of Decision Tree Algorithm for Predicting Student's Performance Via Online Learning During Corona Virus Pandemic", *Journal of Theoretical and Applied Information Technology*, Vol. 99, No. 19, 2021, pp. 4546-4556.

[5] F. Aldi, A. A. Rahma, "University Student Satisfaction Analysis on Academic Services by Using Decision Tree C4.5 Algorithm (Case Study: Universitas Putra "YPTK" Padang)", *International Conference Computer Science and Engineering*, Vol. 1339, 2019. pp. 1-11.

[6] S. Rajendran, S. Chamundeswari, A. A. Sinha, "Predicting the Academic Performance of Middle- and High-School Students using Machine Learning Algorithms", *ScienceDirect*, University of Missouri Columbia (United State of America), 2022, pp. 1-15.

[7] A. AlGhamdi, A. Barsheed, H. AIMshjary, H. AlGhamdi, "A Machine Learning Approach for Graduate Admission Prediction", *Journal Association for Computing Machinery*, King Abdulaziz University (Kingdom of Saudi Arabia), 2020, pp. 155-158.

[8] I. E. Guabassi, Z. Bousalem, R. Marah, A. Qazdar, "A Recommender System for Predicting Students' Admission to Graduate Program using Machine Learning Algorithms", *International Journal of Online and Biomedical Engineering*, Vol. 17, No. 2, 2021, pp. 135-147.

[9] J. Y. Chung, S. Lee, "Dropout Early Warning System for High Schools Students using Machine Learning", *ScienceDirect*, Vol. 96 (C), 2019, pp. 346-353.
[10] M. Nachouki, A. M. Naaj, "Predicting Student Performance to Improve Academic Advising Using the Random Forest Algorithm", *International Journal of Distance Education Technologies*, Vol. 20, No. 1, 2022, pp. 1-17.
[11] A. Asselman, M. Khaldi, S. Aammou, "Enhancing the Prediction of Student Performance Based on the Machine Learning XGBoost Algorithm", *Interactive Learning Environments*, Abdelmalek Essadi University (Marocco), 2021, pp. 1-19.
[12] L. Guang-yu, H. Geng, "The Behavior Analysis and Achievement Prediction Research of College Students Based on XGBoost Gradient Lifting Decision Tree Algorithm", *International Conference of Information and Education Technology*, Beihang University (China), March 29-31, 2019, pp. 289-294.
[13] I. Nirmala, H. Wijayanto, K. A. Notodiputro, "Prediction of Undergraduate Student's Study Completion Status using MissForest Imputation in Random Forest and XGBoost Models", *ComTech Computer, Mathematics and Engineering Applications*, Vol. 13, No. 1, 2022, pp. 53-62.
[14] D. Nashine, "Machine Learning Approach – A Science to Make System Smart – Literature Review", *Proceedings of International Conference on Advances in Computer Technology and Management (ICACTM)*, D. Y. Patil Institute of Master of Computer Applications (India), February 23-24, 2018, pp. 109-112.
[15] D. Papakyriakou, I. S. Barbounakis, "Data Mining Methods: A Review", *International Journal of Computer Application*, Vol. 183, No. 48, 2022, pp. 5-19.
[16] U. Fayyad, G. P. Shapiro, P. Smyth, "From Data Mining to Knowledge Discovery in Databases", *AI Magazine*, Vol. 17, No. 3, 1996, pp. 37-54.
[17] F. Gorunescu, "Data Mining: Concepts, Models, and Technique", Springer-Verlag Berlin Heidelberg, Vol. 12, 2011.
[18] J. W. Li, "Research and Application of Credit Score Based on Decision Tree", *Applied Informatics and Communcation – International Conference ICAIC*, The University of Zhejiang Gongshang (China), August, 2011, pp. 493-501.
[19] L. Breiman, "Random Forest", *Machine Learning*, Vol. 45, No.1, 2001, pp. 5-32.
[20] A. Liaw, M. Wiener, "Classification and Regression by Random Forest", *R News*, Vol. 2, No. 3, 2002, pp. 18-22.
[21] J. H. Freidman, "Greedy Function Approximation: A Gradient Boosting Machine", The Annals of Statistics, Vol. 29, No. 5, 2001, pp. 1189-1232.
[22] T. Chen, C. Guestrin, "XGBoost: A Scalable Tree Boosting System", *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining – KDD '16*, 2016, pp. 785-794.