JITET (Jurnal Informatika dan Teknik Elektro Terapan)
Vol. 11 No. 3, pISSN: 2303-0577 eISSN: 2830-7062 DOI: https://doi.org/10.23960/ jitet.v11i3.3056

# 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, Bina Nusantara University, Jakarta, Indonesia 11480

***Riwayat artikel:***
*Received: 23 Mei 2023*
*Accepted: 10 Juli 2023*
*Published: 1 Agustus 2023*

**Keywords:**
*Graduation Prediction*;
*Students*;
*Machine Learning*;
*Data Mining*.

**Corespondent Email:**

**Abstrak.** 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. Penelitian ini memperoleh data kelulusan siswa dari sekolah yang telah direkapitulasi. Algoritma klasifikasi yang digunakan untuk memprediksi kelulusan siswa 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%.

**Abstract.** *In recent years, there has been a number of students who graduated late at Lubuk Alung 1st State Junior Highschool. This statement is supported by graduation data from Lubuk Alung 1st Satet Junior Highschool. Therefore, it is necessary to predict students' graduation status to identify which factors influence the student's graduation, which will also consequently help the school to solve problem more easily. To solve this problem, the 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 the student's parents. This research retrieved data on student graduation from schools that have been recapitulated. The classification algorithms used to predict students' graduation are decision tree, random forest, and extreme gradient boosting with grid searchCV and k-fold=5. The prediction accuracy using the random forest algorithm outperforms the others with a value of 99.5%*.

## 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 yang menyangkut siswa, namun juga terdapat pola data 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 pembelajaran tambahan.

Lama waktu belajar menjadi salah satu indikator yang penting dalam sistem pendidikan di Indonesia. Lama waktu belajar merupakan durasi studi yang dihabiskan oleh siswa untuk menyelesaikan pendidikan pada jenjang tertentu. 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 yang menyelenggarakan 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]. Algoritma *machine learning* ini merupakan kebutuhan yang dipelajari oleh *machine*. Pembangkit algoritma pembelajaran meliputi proses simulasi pembelajaran cara berpikir manusia, proses penalaran informasi yang tidak lengkap, proses mengkontruksi penemuan hal baru, dan proses pengolahan tren data [2].

Telah dilakukan beberapa penelitian menggunakan *machine learning* yang diterapkan dalam dunia pendidikan. Dilakukan klasifikasi dan prediksi performa akademik mahasiswa menggunakan data yang dibangun dari kehadiran mahasiswa, penilaian praktis, kumpulan nilai tugas, dan skor ujian mahasiswa [3]. Algoritma yang diterapkan adalah *decision tree* dan menghasilkan nilai evaluasi yang baik dengan nilai akurasi 96%. Selain itu penggunaan *decision tree* juga diterapkan untuk mengevaluasi atribut pembelajaran jarak jauh yang dapat mempengaruhi kinerja mahasiswa [4] dan pengukuran kepuasan mahasiswa terhadap layanan akademik [5]. Hasil yang diperoleh dari dua penelitian terakhir menghasilkan nilai evaluasi yang baik dengan menghasilkan nilai akurasi 95,06% [4] dan 95% [5].

Pada penelitian lain, dilakukan prediksi performa akademik siswa sekolah menengah pertama dan siswa sekolah menengah atas menggunakan algoritma *machine learning* berdasarkan beberapa aspek yaitu aspek sosio-demografis, aspek akademik, dan aspek pribadi yang terkait dengan siswa [6]. Penelitian ini menghasilkan nilai evaluasi terbaik pada algoritma *random forest* dengan menghasilkan nilai *precision* 74,94%, nilai *recall* 75,41%, dan nilai *f1-score* 75,14% [6]. Pendekatan ini juga menemukan korelasi bahwa gaya hidup sangat berhubungan dengan prestasi akademik yang diperoleh [6]. Pendekatan *machine learning* digunakan untuk secara otomatis memprediksi kemungkinan penerimaan mahasiswa pasca sarjana untuk membantu lulusan mengenali dan menargetkan universitas yang paling cocok untuk setiap profil mahasiswa dengan menerapkan tiga algoritma *machine learning* yaitu *linear regression*, *decision tree*, dan *random forest* [7]. Penelitian ini menggunakan *admission prediction dataset* dan menghasilkan nilai evaluasi terbaik pada algoritma *random forest* dengan menghasilkan nilai *root mean square error* 7,2% [7].

*Dataset* yang sama pada penelitian [7] digunakan untuk membuat sistem rekomendasi untuk memprediksi mahasiswa awal masuk universitas dengan tujuan membandingkan dan mengevaluasi algoritma yang digunakan dan menentukan parameter paling berpengaruh pada peluang masuk universitas [8]. Penelitian sistem rekomendasi menghasilkan nilai evaluasi terbaik pada algoritma *random forest* dengan menghasilkan nilai *root mean square error* 5,7% dan menemukan bahwa parameter paling berpengaruh pada peluang masuk mahasiswa adalah *CGPA* [8]. Pemodelan prediksi dilakukan menggunakan *machine learning* untuk mengidentifikasi siswa yang berisiko putus sekolah dan membantu mengevaluasinya [9]. Penelitian ini menerapkan algoritma *random forest* dan menggunakan data siswa sekolah menengah atas nasional untuk penyelenggaran informasi pendidikan yang terhubung melalui *internet* [9]. Penelitian ini menghasilkan nilai akurasi 96,88% [9]. Dengan mengkategorikan dan memeriksa faktor *CGPA*, diusulkan model

prediksi *CGPA* untuk mendeteksi performa akademik yang buruk dengan memprediksi rata-rata nilai kelulusan *CGPA* [10]. Penelitian ini membandingkan hasil pekerjaan mereka dengan *benchmark* dimana memperoleh hasil unggul dengan mencatatkan nilai akurasi 86,88% [10].

Penelitian menggunakan algoritma *extreme gradient boosting* dilakukan untuk meningkatkan akurasi prediksi performa mahasiswa [11]. Prediksi yang dilakukan menggunakan *dataset ASSISTments* 2009-2010 dengan menghasilkan nilai akurasi 78,75%, nilai *precision* 75,12%, nilai *recall* 78,75%, dan nilai *f1-score* 73,48% [11]. Penggunaan algoritma *extreme gradient boosting* juga dilakukan untuk menganalisis situasi kehidupan belajar mahasiswa dan memprediksi prestasi mahasiswa [12] dan memprediksi status penyelesaian mahasiswa yang telah mencapai lama studi maksimal [13]. Kedua penelitian terakhir menghasilkan nilai evaluasi yang baik dengan berturut-turut mencatakan nilai akurasi 73% [12] dan nilai akurasi 94,92% [13].

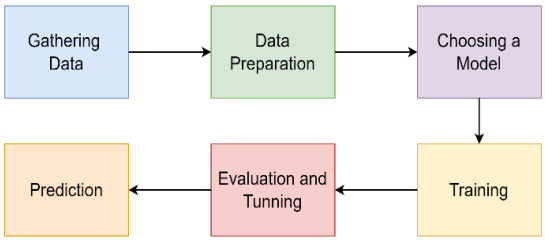
Penelitian yang lebih lanjut dilakukan adalah model prediksi kelulusan siswa sekolah menengah pertama menggunakan algoritma *decision tree*, *random forest*, dan *extreme gradient boosting* untuk meneliti apakah algoritma tersebut mampu melakukan prediksi kelulusan siswa dengan tepat. Algoritma yang digunakan dalam penelitian ini adalah *decision tree*, *random forest*, dan *extreme gradient boosting* yang disesuaikan dengan data kelulusan Sekolah Menengah Pertama Negeri 1 Lubuk Alung. Peneliti akan menginvestigasi algoritma *decision tree*, *random forest*, dan *extreme gradient boosting* untuk konteks prediksi kelulusan siswa serta meneliti optimasi untuk mendapatkan nilai *hyperparameter* yang paling sesuai untuk kasus prediksi kelulusan siswa. Penelitian ini juga akan meneliti variabel yang paling memiliki pengaruh untuk prediksi kelulusan siswa sebagai bahan evaluasi sekolah untuk melakukan perbaikan lebih lanjut di masa yang akan datang.

## 2. TINJAUAN PUSTAKA

### 2.1. *Machine Learning*

*Machine learning* adalah cabang *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.

Algoritma *machine learning* dibagi menjadi *supervised learning*, *unsupervised learning*, dan *reinforcement learning*. Algoritma *supervised learning* umumnya dibagi menjadi regresi dan klasifikasi. Regresi merupakan metode menggunakan fungsi *continue* untuk menyesuaikan dengan variabel *input* dan variabel *output*. Klasifikasi adalah pencocokan variabel *input* dan kategori diskrit. Algoritma *unsupervised learning* merupakan algoritma dimana *output* yang akan dihasilkan tidak diketahui sebelumnya dimana tidak terdapat label atau hanya terdapat satu label dalam *unsupervised learning*. Dan algoritma *reinforcement learning* adalah model pembelajaran observasi melalui mekanisme *try* dan *error*. Hasil yang dianggap baik atau tepat pada algoritma *reinforcement learning* akan disimpan dan menjadi penguat pada *data training* selanjutnya.




Gambar 1. Proses *Machine Learning* [14]

Secara luas proses *machine learning* terdiri dari enam proses utama [14] yaitu:

1. *Gathering Data*
   Langkah pengumpulan data menjadi langkah dasar untuk proses *machine learning*. Meskipun merupakan langkah awal, proses ini sangat penting karena kualitas dan kuantitas data dari proses ini akan membantu menentukan model prediksi.
2. *Data Preparation*
   Setelah data terkumpul dari sumber, selanjutnya mempersiapkan data sehingga dapat digunakan untuk proses training *machine learning*. Visualisasi yang relevan pada data dapat dilakukan untuk menemukan hubungan yang relevan antar

variabel yang berbeda atau menemukan ketidakseimbangan data.

3. *Choosing a Model*
   Proses selanjutnya adalah memilih model yang relevan dengan studi penelitian. Model umumnya dipilih berdasarkan relevansinya terhadap kasus penelitian.
4. *Training*
   Salah satu proses utama dari *machine learning* adalah *training*. Pada proses ini digunakan data dalam perkembangan untuk meningkatkan kemampuan model untuk memprediksi.
5. *Evaluation* dan *Tuning Parameter*
   Setelah proses *training* selesai dilakukan, selanjutnya dilakukan evaluasi untuk memeriksa keakuratannya. Evaluasi akan menggunakan *data test* yang belum pernah digunakan untuk memungkinkan dalam melihat bagaimana kinerja model terhadap data yang belum digunakan. Setelah proses evaluasi, biasanya dilakukan *tuning parameter* untuk melihat apakah berdasarkan parameter yang dimodifikasi akan menyebabkan peningkatan atau penurunan pada hasil evaluasi.
6. *Prediction*
   Proses terakhir dari *machine learning* adalah menggunakan data untuk memberikan jawaban terhadap pertanyaan. Prediksi merupakan proses tujuan untuk menjawab beberapa pertanyaan.

### 2.2. *Data Mining*

*Data mining* merupakan 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* adalah teknik yang digunakan untuk menemukan pengetahuan yang tersembunyi dan hubungan yang tidak terduga antara data. Teknik ini banyak diterapkan dalam berbagai bidang seperti *marketing*, *product development*, *fraud detection*, dan pendidikan (*education*). *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 dari data *input* 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 mengklasifikasikan hasil yang belum diperoleh sebelumnya.

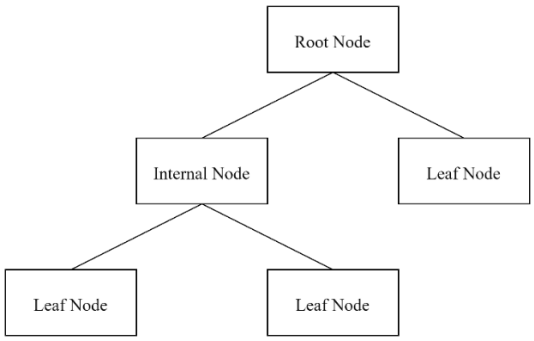
Proses klasifikasi data dibagi dalam dua langkah proses. Proses pertama adalah proses *learning* yang sering disebut dengan fase *training* dimana pada fase ini algoritma klasifikasi membentu aturan klasifikasi data. Proses kedua adalah klasifikasi dimana *data test* digunakan untuk memperkirakan akurasi dari aturan klasifikasi. 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* 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

dimiliki oleh suatu data baru tertentu. Cara kerja *decision tree* dapat dilihat pada Gambar 2.



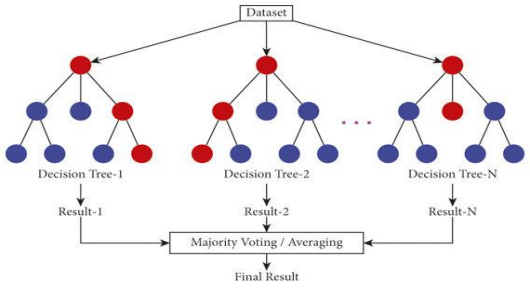

Gambar 2. Cara Kerja *Decision Tree*

## 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 setiap *decision tree* telah dilakukan *training* menggunakan sampel individu dan setiap atribut dipecah pada *tree* yang dipilih antara atribut *subset* yang bersifat acak (*random*). Pada proses klasifikasi, individunya didasarkan pada *vote* dari suara terbanyak pada kumpulan populasi *tree*.

*Random forest* merupakan pengembangan dari metode *CART* yaitu dengan menerapkan metode *bootstrap aggregating* (*bagging*) dan *random feature selection* [19]. *Random forest* memiliki beberapa kelebihan yaitu dapat meningkatkan hasil akurasi jika terdapat data yang hilang dan untuk *resisting outliers*, serta efisien untuk penyimpanan sebuah data. *Random forest* juga mempunyai seleksi fitur dimana mampu mengambil fitur terbaik sehingga dapat meningkatkan performa terhadap model klasifikasi. Dengan adanya seleksi fitur, *random forest* dapat bekerja pada data dengan parameter yang kompleks secara efektif. Dalam *random forest*, banyak pohon yang ditumbuhkan sehingga terbentuk hutan (*forest*), kemudian analisis dilakukan pada kumpulan pohon.

*Random forest* dikembangkan dengan ide bahwa perlu terdapat penambahan *layer* pada proses *resampling* acak pada *bagging*. Selain itu data sampel diambil secara acak untuk membentuk *decision tree*, variabel prediktor juga diambil sebagian secara acak dan baru dipilih sebagai pemilah terbaik saat penentuan pemilah pohon [20].

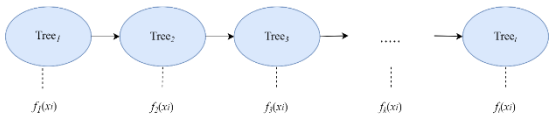



Gambar 3. Cara Kerja *Random Forest* [19]

## 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 *gradient boosting* 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 membangun model baru untuk memprediksi *error* dari model sebelumnya. Pohon pertama dalam *xgboost* cenderung lemah dalam melakukan klasifikasi. Penambahan pohon dilakukan dengan tujuan tidak terdapat lagi perbaikan *error* yang dapat dilakukan. Cara kerja *xgboost* dapat dilihat pada Gambar 4.



Gambar 4. Cara Kerja *Xgboost*

*Xgboost* menggunakan model yang teratur untuk membangun struktur pohon regresi, sehingga memiliki keunggulan dimana memberikan kinerja yang lebih baik dan mampu mengurangi kompleksitas model untuk menghindari *overfitting*. Hasil prediksi akhir *xgboost* biasanya berupa penjumlahan hasil prediksi dari setiap pohon regresi [22]. Pada algoritma *xgboost*, penentuan jumlah pohon dan

*depth* merupakan salah satu indikator utama. 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. PENELITIAN TERKAIT

Penelitian dalam dunia pendidikan menggunakan *machine learning* sudah cukup sering dilakukan. Penelitian juga dilakukan dengan mengelompokkan atau mengklasifikasikan siswa yang lulus tepat waktu dan terlambat. Beberapa algoritma *machine learning* seperti *decision tree*, *random forest*, dan *extreme gradient boosting* memberikan hasil yang cukup optimal seperti yang terlihat pada Tabel 1.

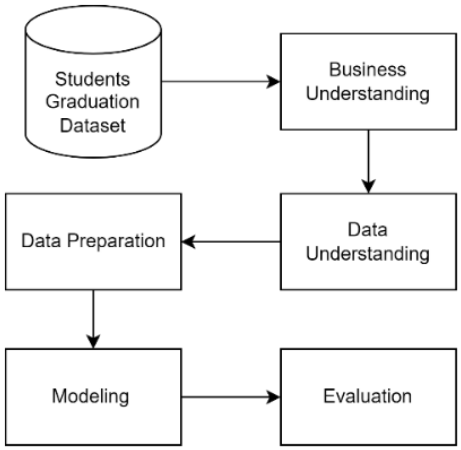
Tabel 1. Penelitian Terkait

| No | Judul | Algoritma | Hasil |
|---|---|---|---|
| 1 | *Decision Tree Algorithms Use in Predicting Students' Academic Performance in Advanced Programming Course* [3] | *Decision Tree* | *Accuracy*: 96% |
| 2 | *Application of Decision Tree Algorithm for Predicting Students' Performance Via Online Learning During Corona Virus Pandemic* [4] | *Decision Tree* | *Accuracy*: 95,06% |
| 3 | *University Student Satisfaction Analysis on Academic Service by Using Decision Tree C4.5 Algorithm* (*Case Study*: Universitas Putra Indonesia "YPTK" Padang [5] | *Decision Tree* | *Accuracy*: 95% |
| 4 | *Predicting the academic performance of middle-and high-school students using machine learning algorithms* [6] | *Random Forest* | *Precision*: 74,94% *Recall*: 75,41% *F1-Score*: 75,14% |
| 5 | *A Machine Learning Approach for Graduate Admission Prediction* [7] | *Random Forest* | *Root Mean Square Error*: 7,2% |
| 6 | *A Recommender System for Predicting Students Admission to a Graduate Program using Machine Learning Algorithms* [8] | *Random Forest* | *Root Mean Square Error*: 5,7% |
| 7 | *Dropout Early Warning Systems for High School Students Using Machine Learning* [9] | *Random Forest* | *Accuracy*: 96,88% |
| 8 | *Predicting Students Performance to Improve Academic Advising Using the Random Forest Algorithm* [10] | *Random Forest* | *Accuracy*: 86,88% |
| 9 | *Enhancing the prediction of student performance based on the machine learning XGBoost algorithms* [11] | *Extreme Gradient Boosting* | *Accuracy*: 78,75% *Precision*: 75,12% *Recall*: 78,75% *F1-Score*: 73,48% |
| 10 | *The Behavior Analysis and Achievement Prediction Research of* | *Extreme Gradient Boosting* | *Accuracy*: 73% |

| | | | |
|---|---|---|---|
| | *College Students Based on XGBoost Gradient Lifting Decision Tree Algorithms* [12] | | |
| 11 | *Prediction of Undergraduate Students Study Completion Status Using MissForest Imputation in Random Forest and XGBoost Models* [13] | *Extreme Gradient Boosting* | *Accuracy*: 94,92% *Sensitivity*: 90,21% *G-Mean*: 92,90% *AUC*: 97,77% |

Berdasarkan beberapa penelitian yang telah dilakukan sebelumnya di dunia pendidikan, *machine learning* mampu melakukan tugas prediksi klasifikasi dengan baik. Algoritma *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 dalam penelitian ini menunjukkan hasil evaluasi yang baik dimana *decision tree* mampu menghasilkan nilai akurasi yang sangat baik. Dalam penelitian lain, algoritma *random forest* diterapkan untuk memprediksi performa akademik siswa sekolah menengah pertama dan 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 dengan menggunakan *random forest* juga menunjukkan hasil evaluasi yang baik dan mampu mengidentifikasi variabel yang 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 dari penelitian dengan menggunakan algoritma *extreme gradient boosting* juga 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. Penelitian akan menggunakan algoritma *decision tree*, *random forest*, dan *extreme gradient boosting* pada kasus memprediksi kelulusan siswa sekolah menengah pertama menggunakan *dataset* kelulusan Sekolah Menengah Pertama Negeri 1 Lubuk Alung.

## 4. METODE PENELITIAN




Gambar 5. Tahapan 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.

## 5. HASIL DAN PEMBAHASAN

### 5.1. *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 dengan tidak tepat waktu. Oleh karena itu pihak sekolah bermaksud untuk melakukan prediksi kelulusan siswa guna mengetahui penyebab keterlambatan kelulusan

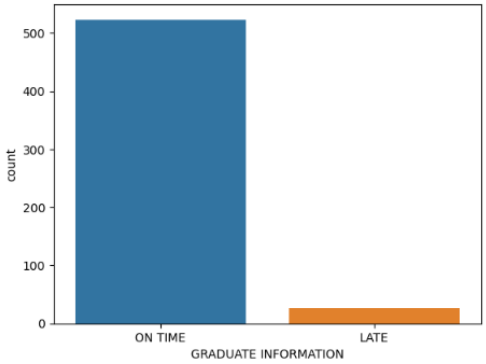
siswa, terutama variabel-variabel yang paling mempengaruhi masalah tersebut menurut data yang tersedia.

### 5.2. *Data Understanding*

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

#### 5.2.1. Eksplorasi Data Kelulusan Siswa

Pertama dilakukan visualisasi berdasarkan jumlah siswa yang lulus tepat waktu dan jumlah siswa yang lulus terlambat. Hasil dari visualisasi dapat dilihat pada Gambar 6.



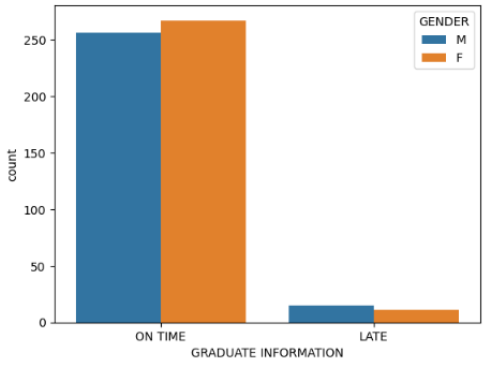

Gambar 6. Jumlah Siswa Berdasarkan Informasi Kelulusan

Berdasarkan visualisasi terlihat bahwa jumlah siswa yang lulus tepat waktu (*on time*) lebih banyak dibandingkan jumlah siswa yang lulus terlambat (*late*). Detail dari jumlahnya dapat dilihat pada Tabel 2.

Tabel 2. Detail Jumlah Siswa Berdasarkan Informasi Kelulusan

| No | *Graduate Information* | Jumlah Siswa |
|---|---|---|
| 1 | *On Time* | 523 |
| 2 | *Late* | 26 |

Selanjutnya dilakukan visualisasi berdasarkan jenis kelamin siswa. Hasil dari visualisasi dapat dilihat pada Gambar 7.



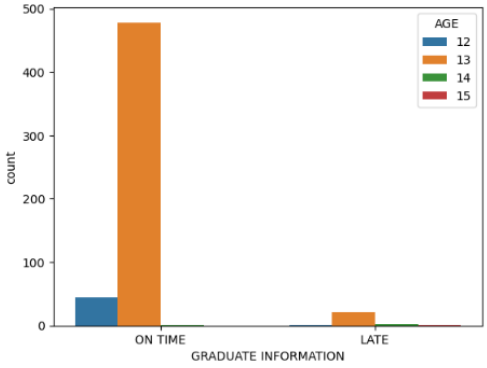

Gambar 7. Jumlah Siswa Berdasarkan Jenis Kelamin Siswa

Berdasarkan visualisasi terlihat bahwa jumlah siswa yang lulus tepat waktu (*on time*) lebih banyak berjenis kelamin perempuan (*female*) daripada laki-laki (*male*). Selain itu terlihat bahwa jumlah siswa yang lulus terlambat (*late*) lebih banyak berjenis kelamin laki-laki daripada perempuan. Detail jumlahnya dapat dilihat pada Tabel 3.

Tabel 3. Detail Jumlah Siswa Berdasarkan Jenis Kelamin Siswa

| No | *Gender* | *Graduate Information* | Jumlah Siswa |
|---|---|---|---|
| 1 | *Male* | *On Time* | 256 |
| | | *Late* | 15 |
| 2 | *Female* | *On Time* | 267 |
| | | *Late* | 11 |

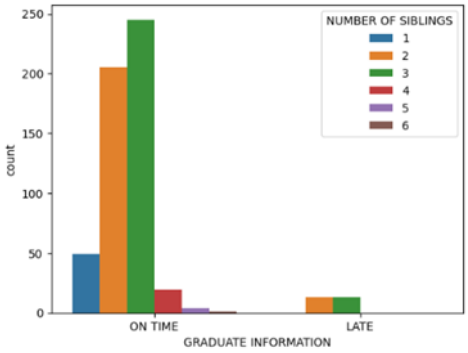
Lalu dilakukan visualisasi berdasarkan usia siswa. Hasil dari visualisasi dapat dilihat pada Gambar 8.




Gambar 8. Jumlah Siswa Berdasarkan Usia Siswa

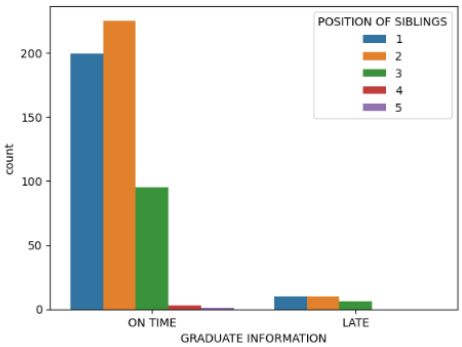
Berdasarkan visualisasi terlihat bahwa jumlah siswa yang siswa yang lulus tepat waktu (*on time*) lebih banyak pada usia 13 tahun saat

awal memasuki sekolah, lalu diikuti oleh usia 12 tahun. Jumlah siswa yang lulus terlambat (*late*) lebih banyak pada usia 13 tahun. Kemudian dilakukan visualisasi berdasarkan jumlah siswa bersaudara. Hasil visualisasi dapat dilihat pada Gambar 9.



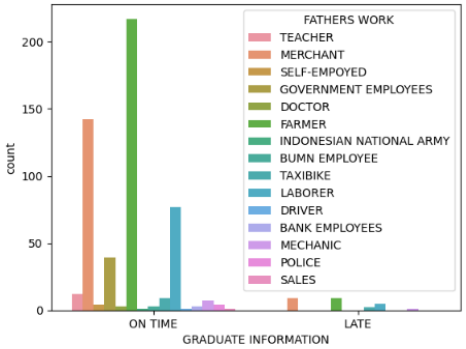

Gambar 9. Jumlah Siswa Berdasarkan Jumlah Siswa Bersaudara

Berdasarkan visualisasi terlihat bahwa jumlah siswa yang lulus tepat waktu (*on time*) lebih banyak dari siswa dengan tiga jumlah bersaudara dan diikuti oleh siswa dengan dua jumlah berasaudara. Jumlah siswa yang lulus terlambat (*late*) lebih banyak pada siswa dengan dua dan tiga jumlah bersaudara. Lalu dilakukan visualisasi berdasarkan posisi siswa dalam jumlah bersaudara. Hasil visualisasi dapat dilihat pada Gambar 10.



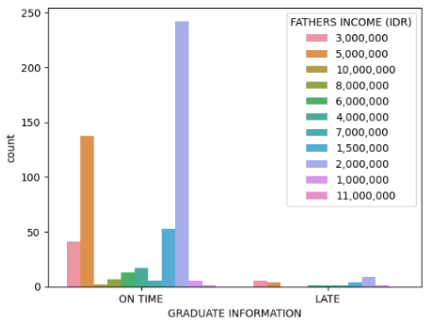

Gambar 10. Jumlah Siswa Berdasarkan Posisi Siswa Dalam Jumlah Bersaudara

Berdasarkan visualisasi terlihat bahwa jumlah siswa yang lulus tepat waktu (*on time*) lebih banyak berada pada siswa anak ke-2 dalam jumlah bersaudara dan diikuti pada siswa anak ke-3 dalam jumlah bersaudara. Jumlah siswa yang lulus terlambat (*late*) lebih banyak berada pada anak ke-1 dalam jumlah bersaudara dan anak ke-2 dalam jumlah bersaudara. Setelah dilakukan eksplorasi berdasarkan data pribadi yang berhubungan dengan siswa, dilakukan eksplorasi berdasarkan data yang berhubungan dengan pekerjaan orang tua siswa. Eksplorasi data pertama dilakukan berdasarkan pekerjaan ayah untuk mengetahui persebaran siswa yang lulus tepat waktu (*on time*) dan siswa yang lulus terlambat (*late*) berdasarkan pekerjaan ayah siswa. Hasil visualisasi dapat dilihat pada Gambar 11.




Gambar 11. Jumlah Siswa Berdasarkan Pekerjaan Ayah Siswa

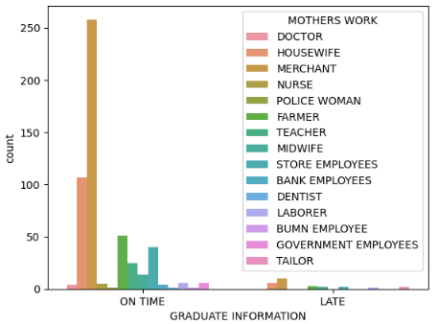
Dari visualisasi jumlah siswa berdasarkan pekerjaan ayah siswa terlihat bahwa sebagian besar ayah siswa memiliki pekerjaan sebagai petani (*farmer*), lalu disusul dengan ayah yang bekerja sebagai pedagang (*merchant*). Siswa yang lulus terlambat (*late*) banyak berada pada ayah memiliki pekerjaan sebagai petani dan pedagang. Kemudian dilakukan eksplorasi berdasarkan penghasilan ayah dari siswa. Hasil visualisasi dapat dilihat pada Gambar 12.




Gambar 12. Jumlah Siswa Berdasarkan Penghasilan Ayah Siswa

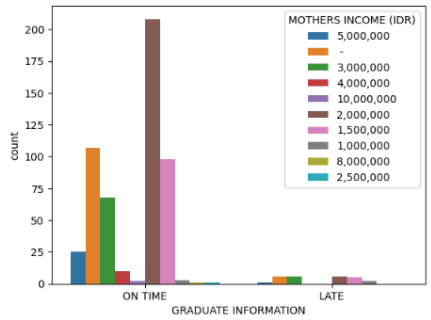
Dari visualisasi jumlah siswa berdasarkan penghasilan ayah siswa terlihat bahwa sebagian besar siswa memiliki ayah dengan penghasilan Rp.2.000.000. Siswa terbanyak yang lulus terlambat (*late*) adalah dengan ayah yang

memiliki penghasilan Rp.2.000.000. Selain data terkait pekerjaan ayah, terdapat juga data terkait pekerjaan ibu siswa. Selanjutnya dilakukan eksplorasi berdasarkan pekerjaan ibu siswa untuk mengetahui persebaran siswa yang lulus tepat waktu (*on time*) dan siswa yang lulus terlambat (*late*) berdasarkan pekerjaan ibu siswa. Hasil dari visualisasi dapat dilihat pada Gambar 13.




Gambar 13. Jumlah Siswa Berdasarkan Pekerjaan Ibu Siswa

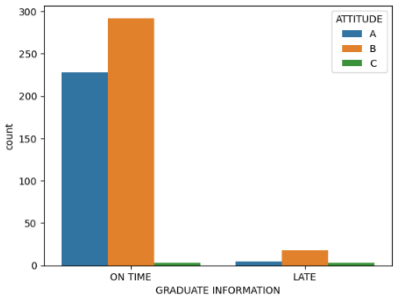
Dari visualisasi jumlah siswa berdasarkan pekerjaan ibu terlihat bahwa sebagian besar siswa memiliki ibu yang memiliki pekerjaan sebagai pedagang (*merchant*). Siswa yang lulus terlambat (*late*) paling banyak memiliki ibu yang berprofesi sebagai pedagang. Kemudian dilakukan eksplorasi berdasarkan penghasilan ibu dari siswa. Hasil visualisasi dapat dilihat pada Gambar 14.




Gambar 14. Jumlah Siswa Berdasarkan Penghasilan Ibu Siswa

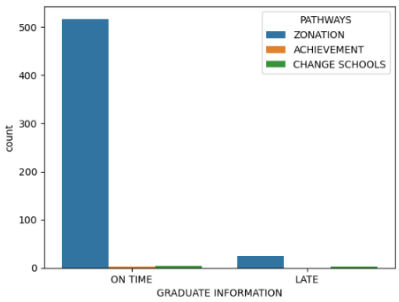
Dari visualisasi jumlah siswa berdasarkan penghasilan ibu siswa terlihat bahwa sebagian besar siswa memiliki ibu dengan penghasilan Rp.2.000.000 dan diikuti oleh ibu dengan penghasilan Rp.3.000.000. Siswa yang paling banyak lulus terlambat (*late*) memiliki ibu yang memiliki penghasilan Rp.2.000.000 dan Rp.3.000.000.

Setelah dilakukan eksplorasi berdasarkan data pribadi yang berhubungan dengan siswa dan pekerjaan orang tua siswa, langkah selanjutnya adalah melakukan eksplorasi berdasarkan data akademik siswa. Data akademik yang digunakan adalah nilai rapor siswa pada semester pertama memasuki jenjang pendidikan sekolah menengah pertama. Eksplorasi tidak dilakukan untuk semua mata pelajaran karena wajib akan dimasukkan ke dalam tahap pemodelan (*modelling*). Oleh karena itu yang menjadi perhatian peneliti adalah menggali data nilai sikap dan jalur masuk siswa di Sekolah Menengah Pertama Negeri 1 Lubuk Alung.




Gambar 15. Jumlah Siswa Berdasarkan Nilai Sikap Siswa

Dari visualisasi jumlah siswa berdasarkan nilai sikap siswa terlihat bahwa sebagian besar siswa memiliki nilai sikap B. Siswa yang lulus terlambat (*late*) sebagian besar memiliki nilai sikap B. Kemudian dilakukan eksplorasi data berdasarkan jalur masuk siswa untuk mengetahui persebaran siswa. Hasil visualisasi dapat dilihat pada Gambar 16.



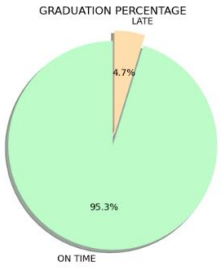

Gambar 16. Jumlah Siswa Berdasarkan Persebaran Jalur Masuk

Dari visualisasi jumlah siswa berdasarkan jalur masuk terlihat bahwa sebagian besar siswa berasal dari jalur masuk zonasi (*zonation*).

Sebagian besar mahasiswa yang lulus terlambat (*late*) berasal dari jalur masuk zonasi.

### 5.2.2. Verifikasi Kualitas Data

Setelah melakukan eksplorasi data yang digunakan dalam penelitian ini, langkah selanjutnya yang dilakukan adalah memverifikasi kualitas data. Tugas pertama dari langkah ini adalah memeriksa apakah terdapat data yang hilang (*missing value*) menggunakan fungsi *isna* dan *isnull*. Berdasarkan hasil pemeriksaan *missing value* pada data yang digunakan, ditemukan bahwa tidak terdapat data yang hilang. Hal ini menunjukkan bahwa kualitas data yang digunakan cukup baik. Selanjutnya, berdasarkan eksplorasi data pada Tabel 2, ditemukan bahwa jumlah data dengan informasi kelulusan tepat waktu (*on time*) lebih banyak dibandingkan dengan informasi kelulusan terlambat (*late*). Lalu dilakukan pemeriksaan keseimbangan data dengan menggunakan visualisasi yang menghasilkan persentase kelulusan siswa berdasarkan informasi kelulusan. Hasil visualisasi dapat dilihat pada Gambar 17.




Gambar 17. Persentase Jumlah Kelulusan Siswa

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

## 5.3. *Data Preparation*

Setelah tahapan eksplorasi data selesai dilakukan secara keseluruhan, maka proses yang dilakukan pada tahap ini terdiri dari mengatasi ketidakseimbangan data (*imbalance data*) dan menghapus variabel yang dianggap tidak relevan berdasarkan proses pemahaman data. Hasil akhir dari proses ini adalah kumpulan fitur yang siap melalui proses pemodelan (*modelling*).

### 5.3.1. Mengatasi Ketidakseimbangan Data

Berdasarkan hasil proses pemahaman data (*data understanding*), ditemukan bahwa data yang digunakan dalam penelitian ini tidak seimbang. Pada langkah ini, proses mengatasi masalah ketidakseimbangan data dilakukan dengan menggunakan metode *random under sampling* dan *random over sampling*. Dalam metode *random under sampling*, beberapa data dihapus dari kelas mayoritas (kelas '*on time*'). Disisi lain, dalam metode *random over sampling*, beberapa data ditambahkan berdasarkan data kelas minoritas (kelas '*late*') hingga kedua kelas data seimbang.

Data yang seimbang akan dihasilkan berdasarkan kedua metode yang digunakan. Namun pada *random under sampling* memiliki kelemahan dimana kemungkinan besar data dengan informasi *classifier* yang akurat akan terhapus. Selain itu pada *random under sampling*, ada kemungkinan sampel yang dipilih merupakan sampel yang bias sehingga tidak akan menjadi representasi populasi yang akurat dan menyebabkan hasil yang tidak akurat dengan *dataset* yang sebenarnya. Disisi lain, *random over sampling* tidak akan menyebabkan hilangnya informasi pengklasifikasi yang akurat, namun juga akan memiliki kemungkinan *overfitting* karena mereplikasi dari kelas minoritas.

Tabel 4. Hasil *Fixing Imbalanced Data*

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

Mengacu pada hasil proses *random sampling* pada Tabel 4, 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 menghilakan pengklasifikasi (*classifier*) yang sangat penting. Untuk alasan ini, data hasil dari proses *random over sampling* akan digunakan untuk proses selanjutnya dalam penelitian ini.

### 5.3.2. Eliminasi Variabel Tidak Relevan

Berdasarkan hasil proses pemahaman data (*data understanding*), ditemukan bahwa variabel nomor (atau atribut 'No') dalam data tidak relevan dengan penelitian ini, sehingga variabel nomor dihilangkan dari data. Kemudian diperlukan ekstraksi data dengan mengubah data menjadi bentuk yang lebih mudah dipahami melalui 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 *matrix correlation* dari seluruh variabel yang digunakan, dihasilkan bahwa tidak ada variabel yang menghasilkan korelasi yang sangat signifikan. Namun terdapat variabel yang menunjukkan korelasi yang lumayan tinggi seperti *mathematics* (matematika), *social sciences* (ilmu pengetahuan sosial), *natural sciences* (ilmu pengetahuan alam), *Indonesian* (bahasa Indonesia), *art and culture* (seni budaya), *physical education* (pendidikan jasmani), dan *enterpreneurship* (kewirausahaan). Selain itu, terdapat variabel yang tidak banyak berpengaruh pada *matrix correlation* seperti *gender* (jenis kelamin). Karena hal tersebut, maka variabel jenis kelamin akan dihapus sebelum proses pemodelan. Pertimbangan utama untuk menghapus variabel yang tidak terlalu berpengaruh agar algoritma *machine learning* dapat bekerja lebih cepat dan mereduksi model yang kompleks sehingga *output* yang dihasilkan mudah untuk diinterpretasikan yang akhirnya akan meningkatkan akurasi pada model yang dibangun. Atribut yang akan digunakan untuk klasifikasi prediksi ditunjukkan pada Tabel 5.

Tabel 5. Atribut Data yang akan 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* |
| 6 | *Natural Sciences* | *Numerical/ Continuous* |
| 7 | *Social Science* | *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* |

### 5.4. *Modelling and 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* [5], *random forest* [9], dan *extreme gradient boosting* [13] menunjukkan hasil yang baik dalam membuat

prediksi dengan menghasilkan nilai akurasi diatas 90%. Pada penelitian ini, sebelumnya akan dilakukan *hyperparameter tuning* menggunakan *grid searchCV* dengan *k-fold*=5 untuk mendapatkan parameter terbaik yang akan diterapkan pada algoritma yang digunakan. Proses ini dilakukan dengan menggunakan data yang telah disiapkan. Hasil dari proses *hyperparameter tuning* algoritma *machine learning* yang digunakan dapat dilihat pada Tabel 6, Tabel 7, dan Tabel 8.

Tabel 6. Hasil *Hyperpamater Tuning* dari *Decision Tree*

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

Tabel 7. Hasil *Hyperpamater Tuning* dari *Random Forest*

| Parameter | *Grid SearchCV Value* | *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 8. Hasil *Hyperpamater Tuning* dari *Extreme Gradient Boosting*

| Parameter | *Grid SearchCV Value* | *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 | 0.25 |
| *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 *hyperparameter* yang diuji. Proses *hyperparameter tuning* ini menggunakan validasi *k-fold*=5 yang digunakan untuk mengevaluasi kinerja model sebanyak 5 iterasi pada proses *grid searchCV* setiap *hyperparameter*. Hasil proses *hyperparameter tuning* selanjutnya digunakan dalam menentukan prediksi klasifikasi.

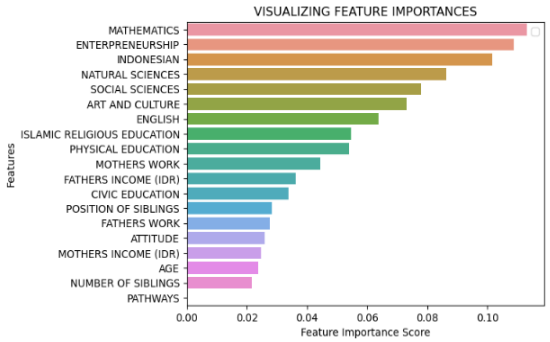
### 5.5. Evaluasi Prediksi Klasifikasi

Pada tahap prediksi klasifikasi, siswa yang lulus tepat waktu (*on time*) dan yang lulus terlambat (*late*) akan diklasifikasikan menggunakan algoritma yang diajukan. Evaluasi yang digunakan adalah dengan menghasilkan nilai akurasi dan nilai *ROC-AUC* dari tiap algoritma. Hasil evaluasi dari prediksi klasifikasi menggunakan ketiga algoritma ditunjukkan pada Tabel 9.

Tabel 9. Hasil Prediksi Klasifikasi

| Algoritma | *Accuracy* | *ROC-AUC* |
|---|---|---|
| *Decision Tree* | 96,66% | 96,56% |
| *Random Forest* | 99,52% | 99,50% |
| *Extreme Gradient Boosting* | 99,04% | 99,01% |

Hasil prediksi klasifikasi kelulusan siswa dengan menggunakan algoritma *random forest* menunjukkan nilai akurasi sebesar 99,5%, sedangkan hasil yang diperoleh menggunakan algoritma *extreme gradient boosting* adalah nilai akurasi sebesar 99,0% dengan skor yang lebih rendah yaitu 0,5%. Hasil prediksi klasifikasi terendah berada pada algoritma *decision tree* yang memiliki akurasi 96,6%. Mengacu pada hasil yang akurat, dapat disimpulkan bahwa variabel yang digunakan dapat memberikan hasil yang memuaskan. 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* dalam memprediksi kelulusan siswa. *Feature importance* menunjukkan hubungan variabel yang digunakan dalam mempengaruhi hasil prediksi kelulusan siswa.




Gambar 18. Hasil Feature Importance

Berdasarkan Gambar 18, nilai matematika (*mathematics*) merupakan variabel yang sangat mempengaruhi kelulusan siswa, diikuti oleh variabel kewirausahaan (*enterpreneurship*) dan bahasa Indonesia (*Indonesian*), serta beberapa variabel lainnya. Oleh karena itu pihak sekolah perlu melakukan evaluasi berupa perbaikan cara memberikan pengajaran pada mata pelajaran matematika, kewirausahaan, dan bahasa Indonesia.

## 6. 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 yang berhubungan dengan pekerjaan orang tua, dan data akademik siswa terbukti cukup efektif dalam memprediksi dan mendeteksi 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 kelulusan siswa yang disebabkan oleh cara mengatasi ketidakseimbangan data karena data yang digunakan sebagian besar merupakan duplikasi dari siswa yang lulus dengan terlambat. Hasil prediksi menunjukkan kecendrungan 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 agar memberikan hasil yang sangat terpercaya. Selain itu, beberapa aspek lainnya yang berkaitan dengan siswa juga dapat diteliti pada penelitian selanjutnya seperti gaya hidup siswa dan jumlah absensi siswa.

## UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Departemen Computer Science Binus, Pihak Sekolah Menengah Pertama Negeri 1 Lubuk Alung, dan rekan-rekan mahasiswa MTI Binus, serta pihak-pihak terkait yang telah memberi dukungan terhadap penelitian ini.

## 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), October 11, 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), October 28, 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), March, 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), May 03, 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.