ISSN 2085-4552

Algoritma Klasifikasi *Decision Tree* C5.0 untuk Memprediksi Performa Akademik Siswa

Natanael Benediktus[1], Raymond Sunardi Oetama[2]
[1, 2] Fakultas Teknik Informatika, Jurusan Sistem Informasi, Universitas Multimedia Nusantara, Tangerang, Indonesia
natanael.benediktus@student.umn.ac.id
raymond@umn.ac.id

Diterima 5 Maret 2020
Disetujui 17 Juni 2020

***Abstract*—Student's performance is often used as a benchmark and a student's activeness is frequently used as a criteria of how well a student academically perform at school. The goal of this study is to find out whether the activeness of a student can predict their academic performance. The data used is an educational dataset is collected using a learning management system (LMS), which is a learner activity tracker tool that is connected by the internet. This data has numerical and categorical variables, so it is needed to have the right algorithm to classify data accurately and ensure data validity. In this study, the C.50 algorithm is used to test the data, where the data is divided into training data by 75% and testing data by 25%. And the result from the tested data, an accuracy of 71.667% is obtained.**

***Index Terms*—C5.0 Algorithm, Data Classification, Decision Tree, Student's Performance**

I. PENDAHULUAN

Performa seorang siswa seringkali disimpulkan dari keaktifannya di dalam ataupun di luar kelas. Siswa yang sering mengangkat tangan dalam kelas atau sering berpartisipasi dalam diskusi, atau yang inisiatif menambah ilmu diluar sekolah, atau bahkan yang jarang absen dalam kelas tentunya tidak jauh dengan label "anak pintar" [1]. Hal seperti ini biasanya hanya ditarik dari sebagian siswa saja yang mendapatkan nilai baik, namun tentunya terdapat sebagian siswa yang memiliki kriteria yang sama, namun memiliki nilai yang kurang memuaskan [2]. Maka dalam penelitian ini ingin dicari tahu apakah benar kriteria yang seringkali digunakan tersebut dapat memprediksi nilai seseorang atau mungkin terdapat kriteria lain yang lebih akurat, bahkan mungkin sebenarnya tidak ada kriteria yang dapat memprediksi nilai seseorang. Perhatian utama dalam penelitian ini adalah untuk mengekstraksi yang selama ini tersembunyi dari data siswa yang dapat meningkatkan performa akademik siswa, karena jika ditinjau banyak sekali aspek dalam dunia ini yang melihat nilai sebagai tolak ukur. Sebagai contoh adalah dalam melamar pekerjaan, sebuah perusahaan tentunya akan melihat nilai dari pelamar kerja sebagai salah satu tolak ukurnya.

Dari beberapa penelitian terdahulu, *Decision Tree* sering dipergunakan untuk predisi prestasi siswa dengan performa yang sangat baik. Algoritma *Decision Tree* untuk memprediksi nilai akhir dari siswa dengan akurasi 85% [3]. Optimasi *Decision Tree* dapat diterapkan untuk prediksi siswa berpotensi bermasalah dengan tingkat akurasi 99,08% [4]. Untuk memprediksi prestasi siswa kelas XII dengan *Decision Tree*, diperoleh hasil akurasi yang diperoleh sangat baik yaitu 97,22% [5]. Sehingga dalam penelitian ini, akan menggunakan metode klasifikasi dengan algoritma *Decision Tree*. Sedangkan beberapa kriteria yang dipergunakan dalam prediksi prestasi siswa juga diambil dari beberapa penelitian terdahulu meliputi: siswa mengangkat tangan [6], berpartisipasi dalam diskusi [7], insisiatif siswa untuk belajar di luar sekolah [8], dan absensi siswa [9].

Berdasarkan latar belakang di atas, dapat ditarik beberapa permasalahan, yang pertama adalah seberapa akuratkah algoritma *Decision Tree* ini dengan faktor keaktifan siswa di dalam kelas maupun di luar kelas yang mencakup keseringan dalam mengangkat tangan, keseringan berpartisipasi dalam diskusi, insiatif siswa untuk belajar di luar sekolah, dan absensi dari siswa itu sendiri untuk memprediksi performa akademik siswa. Yang kedua adalah faktor apa yang paling berpengaruh terhadap performa akademik siswa. Dan yang terakhir adalah kriteria apa yang memperlihatkan performa akademik siswa yang paling baik. Manfaat dari penelitian ini diharapkan dapat memperlihatkan kriteria tertentu yang dapat meningkatkan performa akademik siswa.

II. LANDASAN TEORI

A. *Algoritma C5.0*

Algoritma yang bisa digunakan untuk pembuatan pohon keputusan diantaranya adalah *Decision Tree* [10]. Algoritma C5.0 atau yang lebih dikenal dengan nama *Decision Tree* merupakan salah satu algoritma dari klasifikasi. Algoritma C5.0 ini mengklasifikasikan datanya menjadi pohon bercabang yang memiliki beberapa *node* seperti *root node* yakni *node* di bagian *decision tree* yang paling atas dimana tidak memiliki

panah atau *edge* yang menunjuknya, internal *node* yakni *node* yang memiliki panah atau *edge* yang menunjuk dan juga keluar dari *node* tersebut, dan terminal *node* (*leaf*) yakni *node* di bagian ujung bawah *decision tree* dimana hanya ada panah atau *edge* yang menunjuknya tanpa panah atau *edge* yang keluar dari *node* tersebut [11].

B. Sejarah Algoritma C5.0

Algoritma C5.0 ini merupakan perkembangan dari algoritma C4.5 yang juga merupakan perkembangan dari algoritma ID.3. Algoritma ID.3 (*Iterative Dichotomisher* 3) ini ditemukan oleh John Ross Quinlan sejak tahun 1986. Algoritma ini merupakan metode untuk membangun sebuah *decision tree* yang paling dasar yang melakukan pencarian secara menyeluruh pada semua kemungkinan *decision tree*. Algoritma ID.3 ini lalu dikembangkan menjadi algoritma C4.5 dimana algoritma ini mampu menangani atribut dengan tipe diskrit dan kontinu dimana algoritma ini menggunakan ukuran *entropy* untuk memilih atribut yang tepat. Algoritma C4.5 juga mampu menangani *missing value* dimana dengan memberikan nilai berdasarkan nilai yang paling dominan. Dan algoritma ini juga mampu memangkas *decision tree* dimana digunakan untuk mengatasi masalah *over-fitting* yang terjadi karena adanya data yang tidak relevan di data training, hal ini disebabkan oleh banyaknya cabang di *decision tree* yang lalu dipangkas atau yang lebih sering dikenal dengan istilah *pruning* untuk memperkecil pohon yang akhirnya lebih mudah untuk dipahami. Algoritma C4.5 ini pun dikembangkan menjadi algoritma C5.0 dikarenakan masih terdapat berbagai kelemahan di algoritma C4.5 seperti terjadinya *overlapping* terutama saat data yang dikelola sangat banyak, yang juga menyebabkan meningkatnya waktu pengambilan keputusan. Dengan munculnya algoritma C5.0 ini, tingkat akurasi yang dimiliki lebih tinggi, lebih cepat untuk mengambil keputusan, dan penggunaan memori yang jauh lebih rendah dari algoritma sebelumnya [12].

C. Kegunaan Algoritma C5.0

Algoritma C5.0 digunakan untuk mengklasifikasikan data yang menghasilkan model *decision* keputusan dan aturan-aturan keputusan. Aturan-aturan keputusan ini akan digunakan untuk membantu dalam pengambilan keputusan yang tepat. Algoritma C5.0 juga digunakan untuk mengekstraksi data dimana dapat ditemukan hubungan variabel tertentu dengan variabel target [13].

D. Kelebihan dan Kelemahan Algoritma C5.0

Kelebihan dari algoritma C5.0 adalah kemampuannya untuk menangani masalah seperti *missing value* dan data dengan jumlah yang besar. Algoritma ini juga dapat melakukan *training* data dalam waktu yang cepat untuk digunakan dalam *testing* data. Algoritma C5.0 ini juga menawarkan metode *boosting* yang dapat meningkatkan tingkat akurasi [14].

Kelemahan dari algoritma C5.0 ini sama seperti *decision tree* lainnya yakni model prediksi akan menjadi tidak stabil apabila varians data sangat kecil, dan ketergantungan akan informasi dimana perubahan kecil dalam inputan data dapat menyebabkan perubahan yang besar pada pohon [15]. Sedangkan *Pseudocode* C5.0 [16] dapat dilihat pada Gambar 1.

Pseudo Code:

1. Check for base cases.
2. For each attribute calculate the Normalized information gain for splitting an attribute.
3. Out of this select the best attribute which has the highest information gain.
4. Find a decision node that splits the best, as root node.
5. Recurs on the sub lists obtained by splitting on best of a and add those nodes as children node.

Gambar 1. *Pseudocode* C5.0

III. METODOLOGI

A. Objek Penelitian

Dataset yang digunakan dalam penelitian adalah data dari siswa yang mencakup 3 macam kategori dengan nilai akhir dari siswa. Kategori pertama yaitu fitur demografis seperti jenis kelamin dan kewarganegaraan. Kategori kedua yaitu fitur latar belakang akademik seperti tingkatan pendidikan, dan tingkat kelas dari siswa. Dan kategori ketiga yaitu fitur perilaku siswa seperti berapa kali seorang siswa mengangkat tangan dalam kelas, berapa kali siswa mengunjungi situs tertentu untuk belajar mandiri, berapa kali siswa berpartisipasi dalam diskusi, dan absensi dari siswa tersebut. Dalam penelitian ini, data yang akan digunakan hanyalah fitur perilaku siswa sebagai faktor untuk memprediksi nilai akhir, dimana faktor tersebutlah yang seringkali digunakan sebagai tolak ukur tingginya nilai seorang siswa.

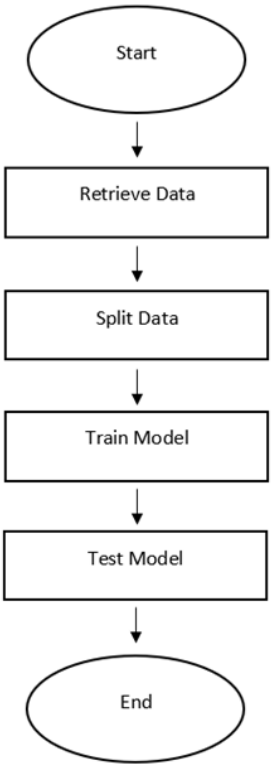
B. Pengumpulan Data

Bentuk dari data yang digunakan dalam penelitian ini berbentuk numerik untuk faktor banyaknya angkat tangan, banyaknya kunjungan siswa dalam situs tertentu untuk belajar mandiri, dan banyaknya partisipasi siswa untuk berdiskusi. Sedangkan, faktor absensi dalam bentuk kategorikal yang dibagi menjadi 2 kategori yaitu lebih dari 7 kali dan kurang dari 7 kali. Dan untuk data nilai akhir siswa dalam bentuk kategorikal juga dimana diabgi menjadi 3 kategori yaitu *high-level* dengan interval nilai 90 sampai dengan 100, *mid-level* dengan interval nilai 70 sampai dengan 89, dan *low-level* dengan interval nilai 0 sampai dengan 69.

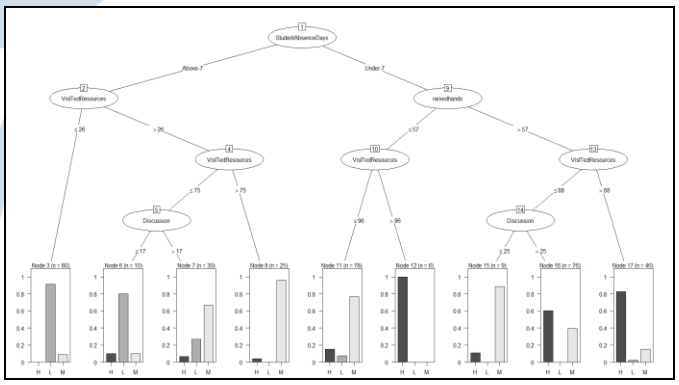
Teknik pengumpulan data yang digunakan adalah dengan teknik observasi dimana data dikumpulkan melalui *Learning Management System* (LMS) yang berfungsi sebagai alat pelacak aktivitas pelajar melalui koneksi internet. Data ini dikumpulkan pada tahun 2016 yang lalu oleh seorang profesor dari sebuah universitas yang bernama "*The University of Jordan*", dimana data ini merupakan data dari siswa sebaran Jordan (https://www.kaggle.com/aljarah/xAPI-Edu-Data). Data ini diyakini sebagai sebuah data yang valid dikarenakan penggunaan *Learning Management System* tersebut dimana observasi tidak hanya sebatas observasi seorang manusia, dimana hal ini tidak mencakup *human error* yang merupakan salah satu faktor akan data yang tidak valid.

Dalam proses *Split* Data, input yang diperlukan berupa dataset dalam RStudio yang merupakan output dari proses "*retrieve data*". Data tersebut akan dibagi menjadi data *training* dan data *testing*, dimana data *training* akan digunakan untuk membangun model, dan data *testing* akan digunakan untuk tes validasi model. Namun, sebelum data dibagi menjadi data *training* dan data *testing*, syntax "set.seed" diperlukan agar nilai akan selalu tetap setiap membagi data.

Total data yang dipergunakan adalah 480. Dalam penelitian ini, data akan dibagi sebesar 75% yaitu 360 untuk data training dan sisanya 25% untuk data testing. Output yang dihasilkan dari proses ini berupa data training dan data testing. Data *training* akan menampung hasil dari temp diatas. Data *testing* akan menampung kebalikannya hasil temp diatas. Dalam proses *Train Model*, input yang diperlukan berupa data yang telah dibagi menjadi data training dan data testing yang merupakan output dari proses "*split data*", dan juga *package* C50 yang akan digunakan untuk membangun model *Decision Tree* C5.0. Output dari proses ini sendiri merupakan *decision tree* dan *decision rule* dari model data training.

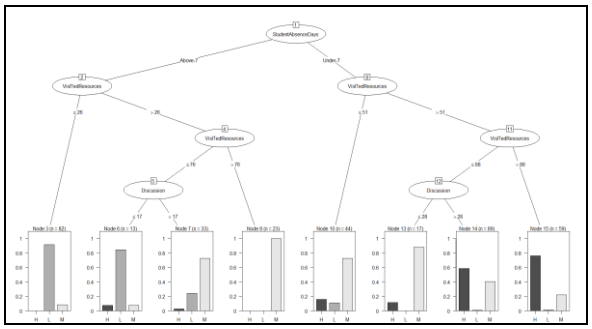


Gambar 2. *Framework*

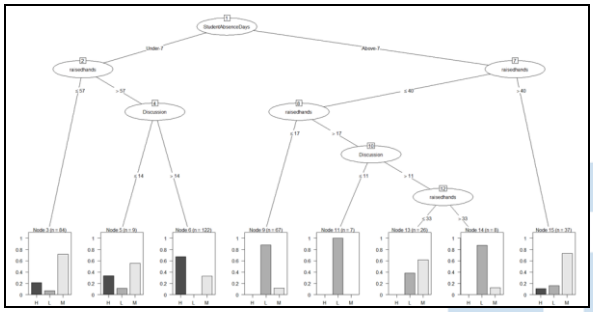


Gambar 3. Model C5.0 seluruh faktor

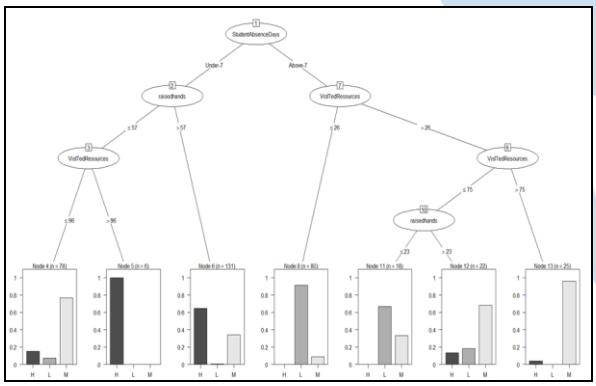
Untuk memvalidasi hasil, akan digunakan tes model dari data *training* yang merupakan output dari proses *train model*, dan data *testing* yang merupakan 25% dari keseluruhan data yang telah dibagi di proses *split* data sebelumnya. Output yang diterima berupa prediksi akan data testing, dimana hasil dari prediksi ini sendiri akan menghasilkan akurasi dari model algoritma C5.0 ini terhadap dataset tersebut. Metode validasi yang digunakan pada penelitian merupakan akurasi dari prediksi model terhadap data tes. Akurasi sendiri merupakan nilai yang mengacu pada tingkat ketepatan prediksi dengan keseluruhan data (Resika Arthana, 2019).



Gambar 4. Model C5.0 tanpa faktor mengangkat tangan



Gambar 5. Model C5.0 tanpa faktor belajar mandiri



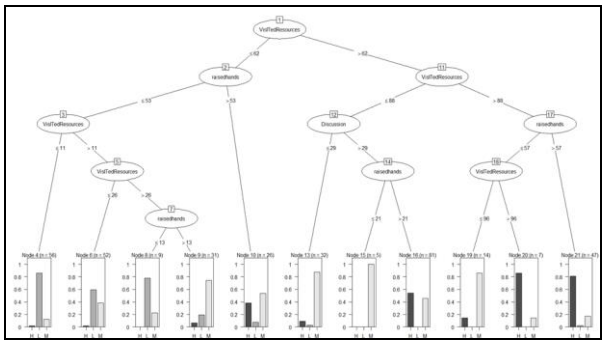
Gambar 6. Model C5.0 tanpa faktor diskusi

IV. HASIL DAN PEMBAHASAN

A. Model Decision Tree pada Semua Faktor

Pada Gambar 3, plot *decision tree* tersebut merupakan plot dari model dengan seluruh faktor. Absensi menjadi faktor yang paling berpengaruh dimana nilai seorang siswa cenderung lebih buruk saat absensinya kurang dari 7 kali, dan nilai seorang siswa cenderung lebih baik saat absensinya lebih dari 7 kali. Faktor selanjutnya yang berpengaruh adalah tingkat keseringan dalam belajar mandiri dan mengangkat tangan

B. Model Decision Tree Tanpa Faktor Angkat Tangan



Gambar 7. Model C5.0 tanpa faktor absensi

Seperti yang dapat dilihat pada gambar 4, plot *decision tree* tersebut merupakan plot dari model tanpa faktor angkat tangan. Absensi menjadi faktor yang paling berpengaruh dimana nilai seorang siswa cenderung lebih buruk saat absensinya kurang dari 7 kali, dan nilai seorang siswa cenderung lebih baik saat absensinya lebih dari 7 kali. Faktor selanjutnya yang menjadi faktor yang berpengaruh adalah tingkat keseringan dalam belajar mandiri.

C. Model Decision Tree Tanpa Faktor Belajar Mandiri

Seperti yang dapat dilihat pada Gambar 5, plot *decision tree* tersebut merupakan plot dari model tanpa faktor belajar mandiri. Absensi menjadi faktor yang paling berpengaruh dimana nilai seorang siswa cenderung lebih buruk saat absensinya kurang dari 7 kali, dan nilai seorang siswa cenderung lebih baik saat absensinya lebih dari 7 kali. Faktor selanjutnya yang menjadi faktor berpengaruh adalah tingkat keseringan dalam mengangkat tangan.

D. Model Decision Tree Tanpa Faktor Diskusi

Seperti yang dapat dilihat pada gambar 6, plot *decision tree* tersebut merupakan plot dari model tanpa faktor diskusi. Absensi menjadi faktor yang paling berpengaruh dimana nilai seorang siswa cenderung lebih buruk saat absensinya kurang dari 7 kali, dan nilai seorang siswa cenderung lebih baik saat absensinya lebih dari 7 kali. Faktor selanjutnya yang menjadi faktor yang berpengaruh adalah tingkat keseringan dalam belajar mandiri dan mengangkat tangan.

E. Model Decision Tree Tanpa Faktor Absensi

Seperti yang dapat dilihat pada gambar 7, plot *decision tree* tersebut merupakan plot dari model tanpa faktor absensi. Faktor yang paling berpengaruh adalah tingkat keseringan dalam belajar mandiri dimana nilai seorang siswa cenderung lebih buruk saat keseringannya dalam belajar mandiri kurang dari

sama dengan 62 kali, dan nilai seorang siswa cenderung lebih baik saat keseringannya dalam belajar mandiri lebih dari 62 kali.

Tabel 1. Validasi hasil

Evaluation Measures	C5.0				
	All Factor	***Without Raised Hands***	***Without Visited Resource***	***Without Discussion***	***Without Absence***
Accuracy	71.67%	74.17%	66.67%	66.67%	57.5%

F. Validasi Hasil

Sebagaimana terlihat pada Tabel 1, akurasi yang terdata merupakan hasil dari *confusion matrix* yang telah dibuat sebelumnya yakni hasil dari proses tes data atas 5 kali percobaan dengan campuran faktor yang berbeda-beda pada algoritma C5.0.

G. Diskusi

Menurut hasil tes dari model yang telah dibuat dari data *training*, akurasi yang dicapai oleh algoritma C5.0 ini mencapai 71.667% untuk prediksi performa akademik siswa dari faktor keaktifan siswa di dalam kelas maupun di luar kelas yang mencakup tingkat keseringan dalam mengangkat tangan, keseringan berpartisipasi dalam diskusi, insiatif siswa untuk belajar di luar sekolah, dan absensi dari siswa itu sendiri pada data testing.

Hasil *modelling* data *training* ini juga memperlihatkan faktor-faktor yang berpengaruh terhadap performa akademik siswa, sebenarnya seluruh faktor sangatlah berpengaruh karena apabila dilihat dari *decision tree* yang telah dilampirkan, tiap kriteria memiliki dampak yang sangat besar terhadap hasilnya. Namun, apabila dilihat secara kesuluruhan absensi seorang siswa dan inisiatif siswa untuk belajar mandiri adalah yang paling berpengaruh terhadap performa akademik mereka, sedangkan partisipasi untuk mengikuti diskusi ataupun tingkat keseringan dalam mengangkat tangan tidak seberpengaruh kedua faktor tersebut.

Dari hasil ini, diperlihatkan juga kriteria terbaik untuk memperoleh performa akademik yang sangat memuaskan, dimana seorang siswa perlu untuk selalu datang ke sekolah dan jarang absen, dimana menurut hasil penelitian ini absensi harus kurang dari 7 kali. Lalu siswa tersebut pun juga harus sering mengangkat tangan dimana mungkin siswa yang jarang untuk mengangkat tangan tentunya memendam rasa keingintahuannya yang malah menjadi tidak mengerti akan pelajarannya, dari hasil penelitian ini menganjurkan untuk mengangkat tangan setidaknya lebih dari 57 kali dalam setahunnya. Lalu, setelah jarang absen dan sering mengangkat tangan, siswa tersebut pun harus sering belajar mandiri, mungkin di situs tertentu, dimana dianjurkan untuk belajar mandiri setidaknya lebih dari 88 kali dalam setahunnya.

V. SIMPULAN

A. Kesimpulan

Kesimpulan yang dapat ditarik dari penelitian ini adalah penggunaan algoritma *Decision Tree* C5.0 ini cukuplah efektif untuk memprediksi performa akademik siswa dimana tercapai akurasi sebesar 71.667%.

Kesimpulan lainnya yang dapat ditarik dari *decision tree* yang telah ditunjukkan adalah seorang siswa memang diperlukan untuk aktif dan selalu datang ke sekolah agar tidak tertinggal pelajaran untuk mendapatkan nilai yang memuaskan. Akan tetapi, memang ada saja siswa yang tidak aktif namun mempunyai nilai yang memuaskan ataupun sebaliknya yang sangat aktif namun mempunyai nilai yang kurang memuaskan. Namun, penelitian ini telah membuktikan bahwa sebagian besar yang aktif memang dapat mendapatkan nilai yang memuaskan.

B. Saran

Untuk peneliti selanjutnya yang ingin atau sedang mengklasifikasikan data bisa menggunakan algoritma *Decision Tree* C5.0 ini, karena hasil yang didapatkan cukup mudah untuk dimengerti terutama untuk para kaum awam.

C. Limitasi

Limitasi dari penelitian ini adalah tidak adanya perbandingan dengan algoritma lain, dimana tentunya akan menghasilkan informasi yang lebih memuaskan dengan algoritma yang berbeda-beda, dimana dapat memberikan kriteria yang berbeda pula.

UCAPAN TERIMA KASIH

Penelitian ini mendapatkan pendanaan dari Universitas Multimedia Nusantara, Tangerang, Indonesia.

DAFTAR PUSTAKA

[1] E. D. Putra, dan I. R. Panglipur, “Analisis Level Kinerja Practitioner Melalui Aktivitas Belajar Siswa”, Jurnal Pendidikan Matematika (JUDIKA EDUCATION), vol. 2, no. 1, hal. 25-35.

[2] D. A. Nurhidayah, D, “Analisis Faktor Kesulitan Belajar Matematika Siswa Sma Pada Implementasi Kurikulum 2013”. In Seminar Nasional Pendidikan, hal. 804-811, 2015.

[3] P. Guleria, N. Thakur, and M. Sood, “Predicting student performance using decision tree classifiers and information gain,” 2014 International Conference on Parallel, Distributed and Grid Computing, 2014.

[4] H. Noor, “Optimasi Model Klasifikasi C4. 5 Dan Particle Swarm Optimization Untuk Prediksi Siswa Bermasalah”, Technologia: Jurnal Ilmiah, vol. 9, no. 4, hal.228-237, 2018.

[5] M. H. Nasution dan M. I. Sofyan, "Penerapan Metode Decision Tree Untuk Memprediksi Prestasi Siswa Kelas XII Dilihat dari Nilai Akhir Semester di SMK Negeri 1 Selong Tahun Pelajaran 2017/2018". Infotek: Jurnal Informatika dan Teknologi, vol. 3, no. 1, hal. 58-65, 2020.

[6] R. J. Sari, dan A. P. Utomo, "Peningkatan Keaktifan Siswa Dan Hasil Belajar Dengan Menggunakan Metode Pembelajaran Problem Based Learning Pada Siswa Smpn 1 Mayang Kelas Ix". ScienceEdu, hal. 80-85, 2019.

[7] M. I. Ibrahim, A. Dassa, dan M. Dinar, "Pengaruh Model Pembelajaran Kooperatif Tipe Think-Talk-Write (TTW) Terhadap Partisipasi Siswa dan Hasil Belajar Siswa dalam Pelajaran Matematika", Issues in Mathematics Education (IMED), vol.1, no.1, hal.26-32, 2019.

[8] R. Kartikasari, S. Suwarjo, dan S. Siswantoro, "Hubungan Bimbingan Belajar di Luar Sekolah dengan Hasil Belajar Matematika", Jurnal Pedagogi, vol.1, no.6, 2019.

[9] K. Malik dan M. Faid, "Prediksi Prestasi Siswa Smp Nurul Jadid Menggunakan Algoritma C4. 5", Nusantara Journal of Computers and Its Applications (NJCA), vol.1, no.2, 2017.

[10] K. Nurkhasanah, "Perbandingan Hasil Penggunaan Metode Decision Tree Dan Random Tree Pada Data Training Aplikasi Pencarian Tukang", ULTIMA INFOSYS, vol.10, no.2, 2019.

[11] R. S. Oetama, "Enhanching Decision Tree Performance in Credit Risk Classification and Prediction", Ultimatics, June 2015.

[12] J. C. Pal, M. A. Hall, E. Frank, I. H. Witten, "Data Mining: Practical Machine Learning Tools and Techniques", Fourth Edition. New Zealand: Morgan Kaufmann, 2016.

[13] I. Kurniawan, dan R. A., "Penerapan Algoritma C5. 0 Pada Sistem Pendukung Keputusan Kelayakan Penerimaan Beras Masyarakat Miskin", Jurnal Informatika, vol.4, no.2, 2017.

[14] N. Sagala dan N. Tampubolon, "Komparasi Kinerja Algoritma Data Mining pada Dataset Konsumsi Alkohol Siswa. Khazanah Informatika", Jurnal Ilmu Komputer dan Informatika, vol.4, no.2, hal.98-103, 2018.

[15] A. P. Wibawa,M. G. A. Purnama, M. F. Akbar, dan F. A. Dwiyanto, "Metode-metode Klasifikasi". In Prosiding Seminar Ilmu Komputer dan Teknologi Informasi (SAKTI), vol.3, no.1, hal. 134-138, 2018.

[16] S. Hardikar, A. Shrivastava and V. Choudhary, "Comparison betweenID3 and C4.5 in Contrast to IDS", VSRDIJCSIT, vol. 2, no. 7, 2012.