

Penerapan Algoritma C4.5 Untuk Memprediksi Nilai Kelulusan Siswa Sekolah Menengah Berdasarkan Faktor Eksternal

Rizky Haqmanullah Pambudi¹, Budi Darma Setiawan², Indriati³

Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Brawijaya
Email: ¹rizkyhaqmanullah@gmail.com, ²s.budidarma@ub.ac.id, ³indriati.tif@ub.ac.id

Abstrak

Pendidikan dalam kehidupan suatu negara memegang peranan yang sangat penting untuk menjamin kelangsungan hidup negara dan bangsa. Statistik menunjukkan bahwa tingkat pendidikan Portugal berada di peringkat terbawah yang disebabkan banyak siswa yang putus sekolah. Faktor eksternal berpengaruh pada kegagalan siswa dalam menyelesaikan bidang studi khususnya bidang studi matematika. Algoritma C4.5 merupakan salah satu metode *data mining* untuk memprediksi kemampuan siswa dalam menyelesaikan bidang studi dilihat dari faktor eksternal siswa. Algoritma C4.5 digunakan untuk mengetahui tingkat akurasi prediksi kemampuan siswa sekolah menengah. Parameter pemilihan fitur adalah faktor-faktor yang mempengaruhi kemampuan siswa sekolah menengah dalam bidang studi matematika. Hasil pengujian dan analisis menunjukkan bahwa Algoritma *Decision Tree C4.5* akurat diterapkan untuk prediksi nilai akhir siswa sekolah menengah dengan tingkat akurasi 60%.

Kata kunci: *data mining, decision tree, algoritma C4.5, prediksi, siswa sekolah menengah*

Abstract

Education in the life of a country plays a very important role to ensure the survival of the state and nation. Statistics show that Portugal's education level is at the bottom of the list due to many students dropping out of school. External factors affect the failure of students in completing the field of study, especially the field of study of mathematics. Algorithm C4.5 is one method of data mining to predict students' ability in completing the field of study seen from the external factors of students. The C4.5 algorithm is used to find out the accuracy of the prediction ability of high school students. The feature selection parameters are the factors that affect the ability of high school students in the field of mathematics studies. Testing and analysis results show that the Decision Tree C4.5 algorithm is accurately applied to predict the final grade of high school students with a 60% accuracy rate.

Keywords: *data mining, decision tree, algorithm C4.5, prediction, school students*

1. PENDAHULUAN

Pendidikan dalam kehidupan suatu negara memegang peranan yang sangat penting untuk menjamin kelangsungan hidup negara dan bangsa. Berdasarkan penelitian yang telah dilakukan oleh Paulo Cortez dan Silvia bahwa selama satu dekade terakhir tingkat pendidikan di Portugis telah meningkat. Akan tetapi statistik menunjukkan bahwa tingkat pendidikan Portugis menempati peringkat terbawah karena banyaknya siswa putus sekolah. Penyebab siswa putus sekolah di Portugal disebabkan kegagalan siswa menyelesaikan beberapa bidang studi yaitu bidang studi matematika dan Bahasa Portugis (Cortez & Silva, 2008).

Ada banyak cara yang dapat digunakan untuk menganalisis kemampuan siswa sekolah menengah, salah satunya yaitu *data mining*. *Data mining* juga dapat digunakan sebagai prediksi untuk memperkirakan nilai masa mendatang. Dengan menerapkan teknik ini akan dibangun pohon keputusan (*decision tree*) untuk kemungkinan siswa yang dapat menyelesaikan studi dengan baik. Salah satu teknik *data mining decision tree* yang terkenal dan dapat digunakan sebagai prediksi adalah algoritma C4.5. Dimana algoritma C4.5 merupakan algoritma klasifikasi data dengan teknik pohon keputusan yang dapat mengolah data numerik (*kontinyu*) dan diskrit, dapat menangani nilai atribut yang hilang, menghasilkan aturan-aturan yang mudah diinterpretasikan dan tercepat diantara

algoritma-algoritma lain. Sebelumnya telah dilakukan penelitian oleh Paulo Cortez dan Alice Silva (2008)

Penelitian tersebut melakukan prediksi siswa-siswa di Portugal yang menjalani bidang studi matematika dan bahasa portugis dengan cara menentukan faktor-faktor yang dapat mempengaruhi nilai siswa. Pada penelitian tersebut ada 3 metode *data mining* yang digunakan untuk memprediksi kemampuan siswa dengan nilai akurasi yang berbeda, yaitu *Naïve Predictor* (60.5%-78.5%), *Decision tree* (62.9%-76.1%), *Random Forest* (33.5%-36.7%).

Penulis memilih menggunakan algoritma *C4.5* karena dapat melakukan prediksi dengan memberikan tingkat nilai akurasi yang ideal untuk memprediksikan kemampuan siswa.

2. DATA PENELITIAN

Data yang digunakan dalam penelitian ini yaitu kemampuan siswa dari sekolah menengah atas di Portugal yang diambil dari situs ICS UCI education, dengan alamat (<http://archive.ics.uci.edu/ml/datasets/student+performance>). Pada *dataset* kemampuan siswa di Portugal terdapat 13 atribut yaitu *sex* (jenis kelamin), *age* (umur), *Medu* (Pendidikan Ibu), *Fedu* (pendidikan ayah), *Mjob* (pekerjaan ibu), *Fjob* (pekerjaan ayah), *reason* (alasan memilih sekolah), *traveltime* (waktu perjalanan ke sekolah), *studytime* (lama waktu belajar), *failures* (kegagalan pada kelas sebelumnya), *schoolsup* (dukungan sekolah), *health* (kesehatan), *absences* (absensi), *G1* (nilai pertama), *G2* (nilai kedua), dan *G3* (sebagai output) Sedangkan untuk kelas atau *output*-nya, yaitu *Excellent* (A), *Good* (B), *Satisfactory* (C), *Sufficient* (D), *Fail* (E).

3. PENDIDIKAN SISWA

Pendidikan dan pengajaran merupakan suatu proses yang sadar tujuan. Tujuan diartikan sebagai suatu usaha untuk memberikan rumusan hasil yang diharapkan oleh siswa setelah mendapatkan pengalaman belajar (Sardiman, 2004). Siswa dapat dibilang berhasil dilihat dari prestasi yang dicapai sebagai tujuan pengajaran. Dengan prestasi yang tinggi, para siswa mempunyai indikasi pengetahuan yang baik.

Motivasi menjadi salah satu faktor penting yang dapat mempengaruhi prestasi siswa di sekolah. Dengan adanya motivasi, siswa mendapat dukungan mental untuk belajar lebih keras, ulet, tekun dan memiliki konsentrasi

penuh dalam proses pembelajaran. Dorongan motivasi perlu dilakukan dalam upaya pembelajaran di sekolah untuk menunjang prestasi siswa. Penting bagi siswa untuk mengenal prestasi belajarnya, karena dengan mengetahui hasil yang sudah dicapai maka siswa mendapat dorongan untuk meningkatkan prestasi yang telah didapatkan sebelumnya (Soemanto, 2003).

3.1. Motivasi Belajar

Pada dasarnya motivasi adalah suatu usaha yang disadari untuk menggerakkan, mengarahkan dan menjaga tingkah laku seseorang agar dia terdorong untuk bertindak melakukan sesuatu sehingga mencapai hasil atau tujuan tertentu. Motivasi dipandang sebagai dorongan mental yang menggerakkan dan mengarahkan perilaku manusia, termasuk perilaku belajar.

3.2. Prestasi Belajar

Poerwanto (2007) memberikan pengertian prestasi belajar yaitu “hasil yang dicapai oleh seseorang dalam usaha belajar sebagaimana yang dinyatakan dalam raport”. Pembelajaran merupakan suatu upaya untuk membelajarkan siswa. Sedangkan merupakan suatu kegiatan yang menghasilkan kemampuan baru yang bersifat permanen pada diri siswa (Seomanto, 2003). Dengan memandang belajar dan pembelajaran sebagai suatu sistem, maka faktor-faktor yang mempengaruhi belajar dan pembelajaran dapat digambarkan sebagai berikut:

1. Faktor internal yaitu faktor-faktor yang berasal dari dalam diri individu dan dapat mempengaruhi hasil belajar individu. Faktor-faktor internal ini meliputi faktor fisiologis dan faktor psikologis.
2. Faktor eksternal Faktor-faktor yang mempengaruhi hasil studi pada siswa bukan hanya bertumpu pada faktor internal yang ada pada siswa saja akan tetapi dari faktor eksternal yang mendukung siswa tersebut. Berikut ini merupakan faktor eksternal yang dapat mempengaruhi studi siswa: (a)Jenis kelasmin, (b)Umur, (c)Waktu belajar, (d)Gagal pada kelas sebelumnya, (e)support sekolah, (f)support keluarga, (g)tambahan bimbingan, (h)motivasi untuk kuliah, (i)kegiatan setelah sekolah, (j)absensi, (k)nilai, (l)masalah siswa.

4. DATA MINING

Menurut Larose (2005), *data mining* didefinisikan sebagai sebuah proses untuk menemukan hubungan, pola dan trend baru yang bermakna dengan menyaring data yang sangat besar dengan menggunakan teknik pengenalan pola seperti teknik statistic dan matematika.

Tugas-tugas dalam *data mining* secara umum dibagi menjadi dua kategori utama:

1. Prediktif

Tujuan dari tugas prediktif adalah untuk memprediksi nilai dari atribut tertentu berdasarkan pada nilai dari atribut-atribut lain. Atribut yang diprediksi umumnya dikenal sebagai target atau variabel tak bebas, sedangkan atribut-atribut yang digunakan untuk membuat prediksi dikenal sebagai *explonatory* atau variabel bebas.

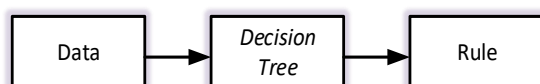
2. Deskriptif

Tujuan dari tugas deskriptif adalah untuk menurunkan pola-pola (korelasi, *trend*, *cluster*, trayektori dan anomali) yang meringkas hubungan yang pokok dalam data. Tugas *data mining* deskriptif sering merupakan penyelidikan dan seringkali memerlukan teknik *postprocessing* untuk validasi dan penjelasan hasil.

5. DECISION TREE

Decision tree adalah teknik model prediksi yang dapat digunakan untuk klasifikasi dan prediksi tugas. *Decision tree* menggunakan teknik “membagi dan menaklukkan” untuk membagi ruang pencarian masalah menjadi himpunan masalah (dunham, 2003).

Proses pada *decision tree* adalah mengubah bentuk data tabel menjadi sebuah model *tree*. Model *tree* akan menghasilkan *rule* dan disederhanakan (Basuki & Syarif, 2003). Konsep pohon keputusan pada Gambar 1.



Gambar 1. Konsep Pohon Keputusan

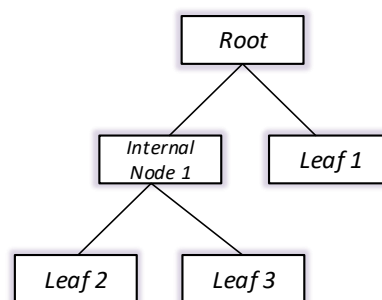
Data dalam *decision tree* biasanya dinyatakan dalam bentuk tabel dengan atribut dan *record*. Atribut menyatakan suatu parameter yang dibuat sebagai kriteria dalam pembentukan *tree*. Pada Gambar 2 adalah contoh penerapan

pohon keputusan dalam memprediksi nilai siswa sekolah.

Nama siswa	Jenis kelamin	Nilai G1	Nilai G2	Nilai Akhir
Rizky	M	12	14	Sangat Baik
Danny	M	11	13	Baik
Made	M	10	12	Cukup
Sarah	F	11	14	Sangat Baik

Gambar 2. Konsep Data Dalam Pohon Keputusan

Decision tree merupakan salah satu teknik klasifikasi terhadap objek atau *record*. Teknik ini terdiri dari kumpulan *decision node*, dan dihubungkan oleh cabang, bergerak ke bawah dari *root node* sampai berakhir di *leaf node* (Yusuf W, 2007). Konsep dasar pohon keputusan pada Gambar 3.



Gambar 3. Konsep Dasar Pohon Keputusan

6. ALGORITMA C4.5

Ada beberapa tahapan dalam membuat sebuah pohon keputusan dalam algoritma C4.5 (Larose, 2005), yaitu:

1. Mempersiapkan data *training*. Data *training* biasanya diambil dari data histori yang pernah terjadi sebelumnya atau disebut data masa lalu dan sudah dikelompokkan dalam kelas-kelas tertentu.
2. Menghitung akar dari pohon. Akar akan diambil dari atribut yang akan terpilih, dengan cara menghitung nilai *gain* dari masing-masing atribut, nilai *gain* yang paling tinggi yang akan menjadi akar pertama. Sebelum menghitung nilai *gain* dari atribut, hitung dahulu nilai *entropy*. Untuk menghitung nilai *entropy* digunakan rumus:

$$Entropy(S) = \sum_{i=1}^n -p_i \log_2(p_i) \quad (1)$$

Keterangan:

S = Himpunan kasus

n = Jumlah partisi S

p_i = Proporsi S_i terhadap S

3. Menghitung nilai *Gain* menggunakan Persamaan 2.

$$Gain(S, A) = entropy(S) - \sum_{i=1}^n \frac{|S_i|}{|S|} Entropy(S_i) \quad (2)$$

Keterangan:

S = Himpunan kasus

A = Fitur

n = Jumlah partisi atribut A

$|S_i|$ = Proporsi S_i terhadap S

$|S|$ = jumlah kasus dalam S

4. Ulangi langkah ke 2 dan langkah ke 3 hingga semua *record* terpartisi
 5. Proses partisi pohon keputusan akan berhenti saat:
 - a. Semua *record* dalam simpul N mendapat kelas yang sama.
 - b. Tidak ada atribut di dalam *record* yang dipartisi lagi.
- Tidak ada *record* di dalam cabang yang kosong.

7. IMPLEMENTASI SISTEM

Aplikasi yang akan dibuat adalah berupa sistem yang dikembangkan untuk melakukan prediksi kemampuan siswa sekolah menengah berdasarkan kategori *excellent*, *good*, *satisfactory*, *sufficient*, *fail* di tiap bidang studi. Input yang digunakan adalah atribut data siswa. Kemudian dilakukan proses pencarian *root* dan pembentukan cabang. Hasil dari proses *training* berupa pohon keputusan yang disimpan dalam bentuk *rule*. *Flowchart* dari proses *training* (pembelajaran) *C4.5* pada Gambar 4.

Pada Gambar 4 terlihat sistem ini memiliki 2 proses utama, yaitu:

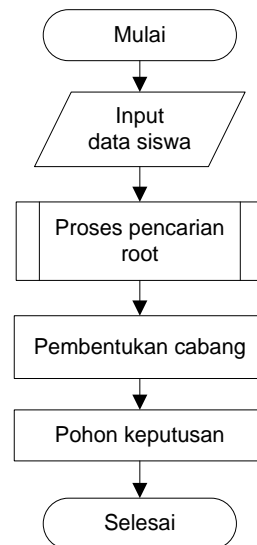
1. Proses pencarian *ROOT*

Proses pencarian *root* dilakukan pertama kali adalah menghitung nilai *entropy* masing-masing kategori. Selanjutnya dilakukan perhitungan *gain* untuk mengetahui nilai *root*. Nilai *gain* tertinggi pada masing-masing kategori akan dijadikan nilai *root*.

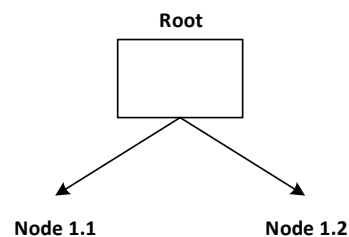
2. Pembentukan Cabang

Setelah ditentukan nilai *root*, maka kategori dengan nilai *gain* tertinggi (*root*) akan dijadikan dasar untuk menentukan pembentukan *Node* (cabang). Kategori dengan nilai *gain* tertinggi

akan dijadikan *Node* selanjutnya, kemudian akan dibentuk pohon *Node* seperti pada Gambar 5.

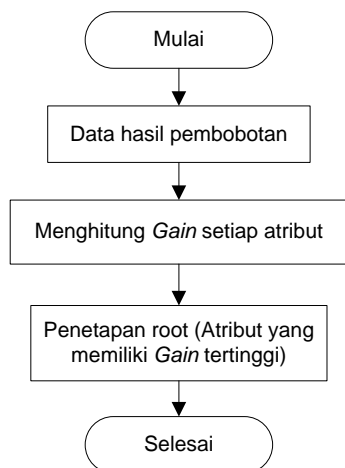


Gambar 4. Flowchart Proses Training (Pembelajaran) *C4.5*



Gambar 5. Pembentukan *Node*

Proses awal dari *training* (pembelajaran) menggunakan algoritma *C4.5* adalah pencarian *root*. Data masukkan pada proses ini berasal dari hasil pembobotan. Kemudian dihitung *information gain* (IG) setiap atribut dimana atribut yang memiliki nilai *information gain* (IG) tertinggi ditetapkan sebagai *root*. Hasil dari proses pencarian *root* adalah satu atribut sebagai *root*. *Flowchart* dari proses pencarian *root* pada Gambar 6.

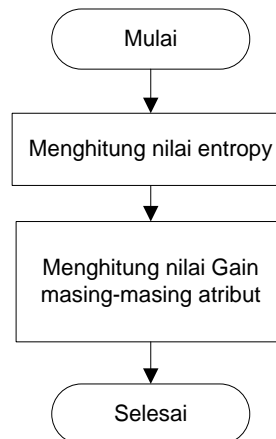


Gambar 6. Flowchart Proses Pencarian Root

Di dalam proses pencarian *root* pada Gambar 6 terdapat proses perhitungan entropy dan perhitungan *information gain* (IG). Mendapatkan nilai entropy dimulai dengan menghitung keseluruhan total kasus *excellent*, *good*, *satisfactory*, *sufficient*, *fail* pada tiap bidang studi. Kemudian menghitung nilai gain masing-masing atribut. Hasil dari proses tersebut diperoleh nilai *information gain* (IG) dari setiap atribut. Flowchart dari proses perhitungan *information gain* (IG) pada Gambar 7.

Proses selanjutnya setelah pencarian *root* adalah proses pembentukan cabang. Dari hasil penentuan *root*, selanjutnya dilakukan kembali proses perhitungan nilai *entropi* dan *gain* untuk menentukan simpul selanjutnya. Data masukkan pada proses ini adalah atribut yang telah terpilih menjadi *root*. Untuk menentukan simpul selanjutnya yaitu dengan menghitung nilai *entropy* dan *gain* semua atribut berdasarkan *root* yang didapat sebelumnya. Proses tersebut akan berjalan rekursif untuk menemukan *node*

selanjutnya dan akan berhenti jika atribut sudah berada pada kelas yang sama atau *entropy* kelas bernilai nol dan pada kondisi tersebut *leaf* akan terbentuk. Hasil dari proses ini adalah pembentukan *decision tree*.

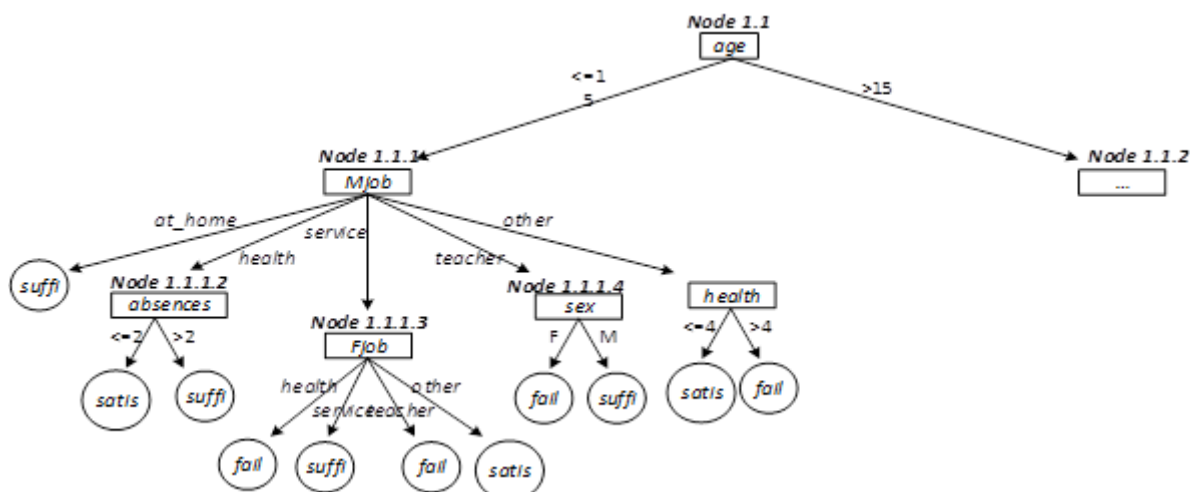


Gambar 7. Flowchart Proses Perhitungan Information Gain

Contoh sebagian hasil perancangan *tree* algoritma C4.5 untuk prediksi kemampuan siswa sekolah menengah dari pada Gambar 8.

Setelah *tree* terbentuk, selanjutnya dilakukan perubahan menjadi *rule*. Berikut ini merupakan pembentukan *rule tree* ditampilkan pada Gambar 8.

1. **IF** atribut *G1* ≤ 11 **AND** atribut *age* ≤ 15 **AND** atribut *Mjob* = *at_home* **THEN** kategori **sufficient**.
2. **IF** atribut *G1* ≤ 11 **AND** atribut *age* ≤ 15 **AND** atribut *Mjob* = *health* **AND** atribut *absences* ≤ 2 **THEN** kategori **satisfactory**.
3. **IF** atribut *G1* ≤ 11 **AND** atribut *age* ≤ 15 **AND** atribut *Mjob* = *health* **AND** atribut *absences* > 2 **THEN** kategori **sufficient**.



Gambar 8. Hasil Pembentukan Tree Siswa Sekolah Menengah

4. IF atribut *G1* <= 11 AND atribut *age* <= 15 AND atribut *Mjob* = *service* AND atribut *Fjob* = *health* THEN kategori *fail*.
5. IF atribut *G1* <= 11 AND atribut *age* <= 15 AND atribut *Mjob* = *service* AND atribut *Fjob* = *service* THEN kategori *sufficient*.
6. IF atribut *G1* <= 11 AND atribut *age* <= 15 AND atribut *Mjob* = *service* AND atribut *Fjob* = *teacher* THEN kategori *fail*.
7. IF atribut *G1* <= 11 AND atribut *age* <= 15 AND atribut *Mjob* = *service* AND atribut *Fjob* = *other* THEN kategori *satisfactory*.
8. IF atribut *G1* <= 11 AND atribut *age* <= 15 AND atribut *Mjob* = *teacher* AND atribut *sex* = F THEN kategori *fail*.
9. IF atribut *G1* <= 11 AND atribut *age* <= 15 AND atribut *Mjob* = *teacher* AND atribut *sex* = M THEN kategori *sufficient*.
10. IF atribut *G1* <= 11 AND atribut *age* <= 15 AND atribut *Mjob* = *other* AND atribut *health* <= 4 THEN kategori *satisfactory*.
11. IF atribut *G1* <= 11 AND atribut *age* <= 15 AND atribut *Mjob* = *other* AND atribut *health* > 4 THEN kategori *fail*.
12. IF atribut *G1* <= 11 AND atribut *age* > 15 AND atribut *reason* = *course* THEN kategori *fail*.
13. IF atribut *G1* <= 11 AND atribut *age* > 15 AND atribut *reason* = *home* AND atribut *Fjob* = *service* AND atribut *Mjob* = *service* THEN kategori *fail*.

8. SKENARIO PENGUJIAN

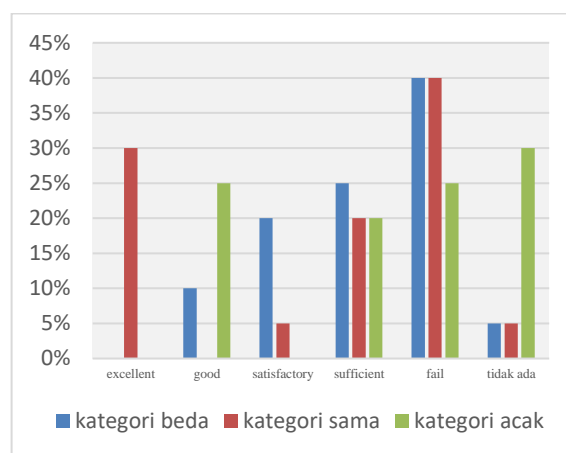
Pengujian ini dilakukan sebanyak 3 kali dengan perbandingan data latih dan data uji yang sama. Data latih pertama menggunakan 60 data latih dengan jumlah tiap kelas *G3* berbeda. Data latih kedua menggunakan 60 data latih dengan jumlah tiap kelas *G3* sama. Data latih ketiga menggunakan 60 data latih secara acak. Data uji yang digunakan berjumlah 20 sampel data siswa. Jumlah data uji dibuat tetap jumlahnya dalam setiap percobaan karena untuk menguji apakah jika menggunakan data latih yang berbeda akan mempengaruhi hasil akurasi prediksi sistem.

8.1 Analisis Hasil

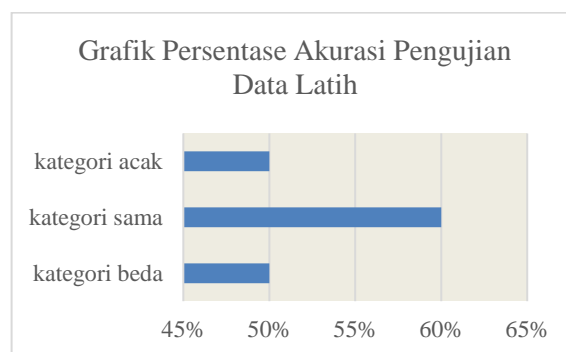
Gambar 9 menunjukkan grafik persentase tiap kategori hasil pengujian data latih. Pada gambar tersebut grafik terlihat hasil prediksi tiap kategori berbeda. Untuk kategori *excellent* dengan persentase tertinggi yaitu 30%, kategori *good* 25%, kategori *satisfactory* 20%, kategori *sufficient* 25%, dan kategori *fail* 40%. Nilai kategori *tidak ada* artinya bahwa *rule* algoritma *C4.5* yang dihasilkan setelah proses pembelajaran tidak mencakup keseluruhan data

uji. Hal ini terjadi karena terdapat *missing value* pada saat proses pembelajaran, sehingga *rule* algoritma *C4.5* yang dihasilkan tidak dapat maksimal. Jumlah kategori prediksi yang digunakan juga mempengaruhi terjadinya *missing value* pada saat proses pembelajaran.

Grafik nilai akurasi prediksi data latih dapat dilihat pada Gambar 10. Nilai akurasi prediksi terbesar terjadi pada data latih dengan jumlah kategori sama dengan persentase 60%. Sedangkan pada data latih dengan jumlah kategori acak dan kategori beda memperoleh nilai akurasi yang sama yaitu 50%. Jumlah kategori berpengaruh besar terhadap tingkat akurasi prediksi algoritma *C4.5*.



Gambar 9. Grafik Kategori Hasil Pengujian Data Latih



Gambar 10. Grafik Perbandingan Nilai Akurasi Prediksi

9. KESIMPULAN

Berdasarkan hasil perancangan, implementasi, dan pengujian yang dilakukan, dapat diambil kesimpulan bahwa Implementasi algoritma *C4.5* untuk prediksi kinerja siswa sekolah menengah dimulai dengan melakukan analisis kebutuhan data kemudian dilakukan penelitian, setelah didapatkan data-data yang diperlukan kemudian dilakukan pembuatan

sistem dengan mengimplementasikan algoritma C4.5 kedalam sistem tersebut dan melakukan pengujian terhadap beberapa sampel data untuk mendapatkan akurasi dari hasil prediksi yang dilakukan oleh sistem yang telah dibuat.

Berdasarkan hasil pengujian akurasi dari pengujian data latih dengan jumlah kategori berbeda menghasilkan tingkat akurasi 50%, pengujian data latih dengan jumlah kategori sama menghasilkan tingkat akurasi 60%, dan pengujian data latih dengan jumlah kategori acak menghasilkan tingkat akurasi 50%.

DAFTAR PUSTAKA

- Basuki, Ahmad dan Syarif, Iwan, 2003. *“Decision Tree”*. Surabaya: Politeknik Elektronika Negeri.
- Cortez, Paulo & Silva, Alice Maria Gonçalves, 2008 *“Using Data mining to Predict Secondary School Student Performance”*. Portugal: University of Minho.
- Dunham, Margareth H., 2003. *“Data Mining Introductory and Advanced Topics”*. New Jersey: Prentice Hall.
- Larose, Daniel T., 2006. *“Data Mining Methods and Models”*. New Jersey: John Willey & Sons, Inc. Hoboken.
- Machine Learning Repository, 2008. *Student Performance Data Set*. Tersedia di <http://archive.ics.uci.edu/ml/datasets/student+performance> [diakses 1 Januari 2017].
- Sardiman, A.M., 2004. *“Interaksi dan Motivasi Belajar Mengajar”*. Jakarta: PT.Raja Grafindo Persada.
- Soemanto, Wasty, 2003. *“Psikologi Pendidikan”*. Jakarta: Rineka Cipta.
- Yusuf W, Yogi, 2007. *“Perbandingan Performasi Algoritma Decision Tree C5.0, CART, dan CHAD: Kasus Prediksi Status Resiko Kredit di Bank X”*. Bandung: Universitas Katolik Parahyangan.