IMPLEMENTASI ALGORITMA *DECISION TREE* UNTUK KLASIFIKASI DATA PESERTA DIDIK

Imam Sutoyo
Fakultas Teknologi Informasi
Universitas Bina Sarana Informatika Jakarta
www.bsi.ac.id
imam.ity@bsi.ac.id

Abstract—*Classification of students aims to classify participants in educational programs so that learning activities can be tailored to these groups. The traditional method for carrying out this classification is by ordering students using single attribute, namely their final value then dividing them according to a certain size. A better method is to use the Data Mining algorithm that is able to use more than one attribute. In this study, the Decision Tree algorithm is used to carry out the classification. The methodology used is CRISP-DM. The Decision Tree algorithms tested are C4.5 and Random Forest. Validation is carried out using 10-Fold Cross Validation to find algorithms that provide the highest precision. Based on the experiment, it was found that Decission Tree C.45 gave the best results with an accuracy of 96,73%. Therefore, in the Deployment phase of the CRISP-DM methodology, the models and rules of C4.5 are used to create applications for this classification.*

Keywords*: Classification Algorithm, Decision Tree, C4.5, Random Forest, CRISP-DM Methodology.*

Intisari—Klasifikasi peserta didik bertujuan untuk mengelompokkan peserta program pendidikan agar kegiatan pembelajaran dapat disesuaikan dengan kelompok-kelompok tersebut. Metode tradisional untuk melaksanakan klasifikasi ini adalah dengan mengurutkan peserta didik menggunakan satu atribut, yaitu nilai akhir mereka kemudian membagi mereka berdasarkan ukuran tertentu. Metode yang lebih baik adalah dengan menggunakan algoritma Data Mining yang mampu menggunakan lebih dari satu atribut. Pada penelitian ini, algoritma *Decision Tree* digunakan untuk melaksanakan klasifikasi. Metodologi yang digunakan adalah CRISP-DM. Algoritma *Decision Tree* yang diujicoba adalah C4.5 dan *Random Forest*. Validasi dilaksanakan menggunakan 10-*Fold Cross Validation* untuk dicari algoritma yang memberikan akurasi paling tinggi. Berdasarkan percobaan, didapatkan hasil bahwasanya *Decission Tree* C.45 memberikan hasil terbaik dengan akurasi 96,73 %. Oleh karena itu, pada tahap *Deployment* dari metodologi CRISP-DM, model dan rule dari C4.5 digunakan untuk membuat aplikasi untuk klasifikasi ini.

Kata Kunci*: Algoritma Klasifikasi, Decision Tree, C4.5, Random Forest, Metodologi CRISP-DM.*

PENDAHULUAN

Pada saat sebuah penyelengara program pendidikan membuka sebuah program pendidikan banyak faktor yang akan mempengaruhi efektifitas program pendidikan tersebut. Efektifitas disini artinya adalah ukuran tercapai atau tidaknya tujuan dari program pendidikan. Beberapa faktor penting yang mempengaruhi efektifitas tersebut selain kurikulum, materi, pengajar, sarana dan prasarana, faktor yang sangat penting lainnya adalah pengelompokkan peserta didik. Penelitian ini bertujuan untuk menghasilkan model klasifikasi untuk keperluan pengelompokkan data peserta didik menggunakan algoritma *Decision Tree*.

Metode yang paling sederhana untuk pengelompokkan ini adalah dengan mengurutkan peserta didik berdasarkan nilai akhir mereka kemudian membagi mereka berdasarkan ukuran tertentu, misalnya jumlah maksimal dalam satu kelas. Jika dibagi menjadi 3 kelas maka kelas pertama yang merupakan kelas atas berisi peserta didik dengan nilai akhir tertinggi, kelas ketiga yang merupakan kelas bawah berisi peserta didik dengan nilai terendah, dan kelas kedua berisi dengan peserta didik dengan nilai diantara kedua kelas atas dan kelas bawah tersebut.

Jadi, metoda klasifikasi sederhana tersebut hanya menggunakan satu attribut saja untuk menentukan seorang peserta didik akan

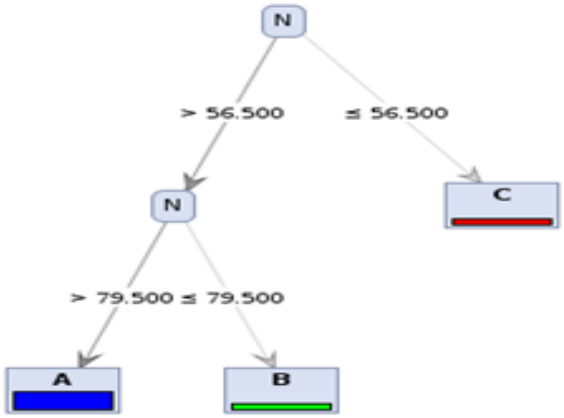
masuk ke kelas yang mana, yakni nilai akhir yang sifat attributnya numerik.

Para peneliti di bidang Data Mining telah menghasilkan banyak algoritma klasifikasi yang dapat digunakan untuk melaksanakan klasifikasi dengan cara yang lebih baik dengan menggunakan beragam attribut, baik sifatnya numerik maupun nominal. Implementasi Data Mining pada bidang pendidikan yang dikenal dengan istilah *Educational Data Mining* (EDM) bertujuan untuk mengembangkan metode yang mampu menemukan pengetahuan-pengetahuan berharga dari data yang dihasilkan pada lingkungan pendidikan (Yadav & Pal, 2012).

Klasifikasi adalah teknik Data Mining yang terbukti paling bermanfaat untuk data di bidang pendidikan (Ahmed & Elaraby, 2014).Telah banyak penelitian Data Mining pada bidang pendidikan khususnya untuk teknik klasifikasi. (Ahmed & Elaraby, 2014) menggunakan algoritma Decision Tree ID3 untuk melaksanakan klasifikasi data peserta didik untuk memprediksi nilai akhir (Yadav & Pal, 2012) menggunakan algoritma Decision Tree ID3, C4.5, dan CART untuk melaksanakan klasifikasi data peserta didik untuk memprediksi hasil ujian. (Saber Iraji, Aboutalebi, Seyedaghaee, & Tosinia, 2012) menggunakan algoritma *Adaptive Neuro Fuzzy* untuk melaksanakan klasifikasi data peserta didik berdasarkan keterkaitan antara faktor-faktor yang mempengaruhi belajar seperti rombongan belajar atau teman-teman sekelas, akses Internet, akses Televisi dan lainnya.

Klasifikasi adalah sebuah proses analisa data yang menghasilkan model-model untuk menggambarkan kelas-kelas yang terkandung dari data tersebut (Han, Kamber, & Pei, 2012). Model - model tersebut disebut *classifier*. Jadi, *classifier* inilah yang akan digunakan untuk menyusun kelas-kelas yang terkandung dari data, misalnya untuk *Decision Tree* maka kelas-kelas tersebut digambarkan dalam bentuk pohon.

Decision Tree digunakan untuk mempelajari klasifikasi dan prediksi pola dari data dan menggambarkan relasi dari variabel attribut x dan variabel target y dalam bentuk pohon (Ye, 2014). *Decission Tree* adalah struktur menyerupai *flowchart* dimana setiap internal node (node yang bukan *leaf* atau bukan node terluar) merupakan pengujian terhadap variabel attribut, tiap cabangnya merupakan hasil dari pengujian tersebut, sedangkan node terluar yakni *leaf* menjadi labelnya (Han et al., 2012). Ada dua jenis algoritma *Decision Tree* yang terkenal, yaitu C4.5 dan *Random Forest*.



Sumber: (Sutoyo, 2018)

Gambar 1. *Decision Tree*

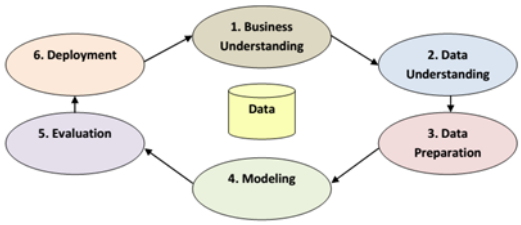
Algoritma C4.5 merupakan algoritma yang dikembangkan dari algoritma ID3. C4.5 ini merupakan algoritma turunan dari algoritma ID3 dengan beragam peningkatan. Beberapa peningkatan ini diantaranya adalah, penanganan atribut-atribut numerik, *missing value* dan *noise* pada dataset, dan aturan-aturan yang dihasilkan dari model pohon yang terbentuk (Larasati & Sutrisno, 2018).

Algoritma *Random Forest* (RF) merupakan algoritma *Decission Tree* yang membentuk model klasifikasi dalam bentuk satu set pohon pada saat proses training dataset. Setiap pohon secara individu bekerja menggunakan beberapa attribut yang dipilih secara acak.

Proses klasifikasi pada RF dilaksanakan dengan mengambil keputusan yang dominan dari setiap pohon yang terbentuk. RF bertujuan untuk menyelesaikan permasalahan *overfitting* pada penggunaan algoritma *Decision tree* (Wambui, George, & Kimani, 2018).

Untuk penelitian Data Mining, telah ada metodologi standar yang disebut CRISP-DM atau *Cross-Industry Standard Process for Data Mining.* CRISP-DM merupakan hasil kolaborasi dari beberapa perusahaan, diantaranya Daimler-Benz, OHRA, NCR Corp., dan SPSS Inc. yang mulai dirintis sejak tahun 1999 (North, 2012).

CRISP-DM memiliki enam tahapan (North, 2012), yaitu:



Sumber: (North, 2012)

Gambar 2. Siklus CRISP-DM

1. *Business Understanding*

Pada tahapan pertama ini harus didefinisikan apa pengetahuan yang ingin didapatkan dalam bentuk pertanyaan-pertanyaan yang sifatnya umum, misalnya bagaimana cara meningkatkan keuntungan, bagaimana cara mengantisipasi kesalahan cacat produk, dan sebagainya.

2. *Data Understanding*

Tahapan kedua ini bertujuan untuk mengumpulkan, mengidentifikasikan, dan memahami aset data yang kita miliki. Data tersebut juga harus dapat diverifikasi kebenaran dan realibilitasnya.

3. *Data Preparation*

Tahapan ini meliputi banyak kegiatan, seperti membersihkan data, memformat ulang data, mengurangi jumlah data, dan sebagainya yang bertujuan untuk menyiapkan data agar konsisten sesuai format yang dibutuhkan.

4. *Modelling*

Model adalah representasi komputasi dari hasil pengamatan yang merupakan hasil dari pencarian dan identifikasi pola-pola yang terkandung pada data.

5. *Evaluation*

Evaluasi bertujuan untuk menentukan nilai kegunaan dari model yang telah berhasil kita buat pada langkah sebelumnya.

6. Deployment

Deployment adalah saat dimana hasil dari seluruh tahapan sebelumnya digunakan secara nyata.

Pada setiap tahapan dari enam tahap CRISP-DM mengandung beberapa pekerjaan. Metodologi CRISP-DM ini sifatnya tidak linear, yakni keenam tahapan tersebut tidak harus dilaksanakan seluruhnya secara berurutan (Cerón, López, & Eskofier, 2018). Meskipun demikian, pada penelitian ini keenam tahapan tersebut dapat dilaksanakan.

Konteks penelitian ini adalah penggunaan metode CRISP-DM dengan algoritma Decision Tree menggunakan dataset yang didapatkan dari hasil pengerjaan quiz secara online oleh peserta didik di tingkat pendidikan tinggi. Data didapatkan dari sistem quiz online dengan menggunakan dua attribut, yaitu nilai rata-rata pengerjaan materi quiz dan jumlah sesi pengerjaan quiz.

BAHAN DAN METODE

Berikut implementasi metodologi CRISP-DM pada penelitian ini.

Business Understanding

Sesuai dengan tahapan CRISP-DM maka pada tahapan pertama, pertanyaan yang bisa dibuat adalah bagaimana cara mengelompokkan siswa dengan efektif berdasarkan atribut-atribut yang berpengaruh? Pengelompokkan ini diharapkan efektif karena lebih komprehensif dimana pengelompokkan dapat dilaksanakan dengan menggunakan lebih dari satu attribut.

Ada dua attribut yang akan digunakan untuk melaksanakan klasifikasi, yaitu nilai dan jumlah sesi pengerjaan. Nilai adalah hasil akhir dari evaluasi sedangkan jumlah pengerjaan adalah hitungan banyaknya evaluasi dikerjakan.

Kedua attribut ini merupakan dua parameter yang tidak bisa ditinggalkan salah satunya untuk klasifikasi siswa secara efektif. Berbeda dengan klasifikasi secara tradisional yang hanya menggunakan nilai saja tanpa memperhitungkan jumlah sesi pengerjaan yang merupakan indikator kesungguhan seorang siswa dalam belajar.

Data Understanding

Dataset yang akan digunakan untuk training adalah data pengerjaan evaluasi secara online. Berikut attribut dari dataset:

1. ID_SISWA: Nomor induk peserta didik, sifatnya numerik, mengidentifikasikan setiap peserta didik secara unik.
2. NAMA_SISWA: Nama peserta didik.
3. JUMLAH_SESI: Jumlah sesi pengerjaan quiz online, sifatnya numerik. Jumlah minimalnya 0 artinya siswa tersebut tidak pernah mengerjakan dan maksimal 50.
4. NILAI_RATA-RATA: Nilai rata-rata dari seluruh sesi pengerjaan quiz, sifatnya numerik. Nilai minimalnya 0 dan maksimalnya 100.
5. KELAS: Kelas dari siswa yang didapatkan dengan mengurutkan siswa berdasarkan 2 variabel, yaitu Nilai Rata-rata dan Jumlah Pengerjaan Quiz. Sifatnya nominal, yaitu A untuk Kelas Atas, B untuk Kelas Menengah ke Atas, dan C untuk Kelas Menengah ke Bawah, dan D untuk Kelas Bawah.

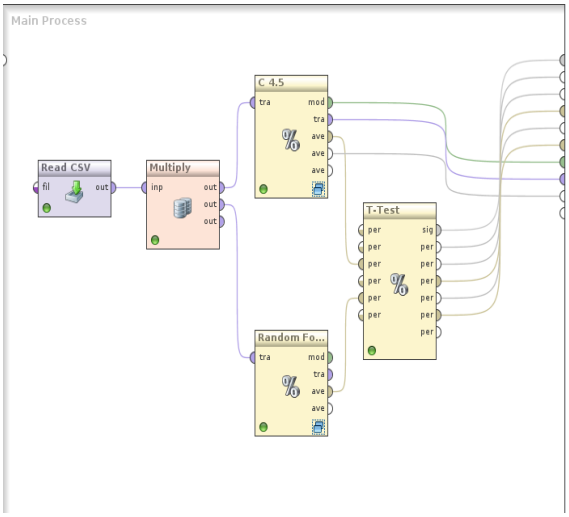
Data Preparation

Tahapan ini sesuai namanya, yakni menyiapkan dataset yang akan menjadi input bagi algoritma klasifikasi. Penyiapan data ini penting terutama untuk menyusun data sesuai format yang sesuai dengan algoritma klasifikasi yang digunakan.

Tahapan ini juga bertujuan untuk menghilangkan data-data yang dapat mengganggu kinerja dari algoritma klasifikasi.

Modelling

Untuk membuat model klasifikasi digunakan Rapidminer. Desain pembentukan model ditunjukkan pada gambar berikut.



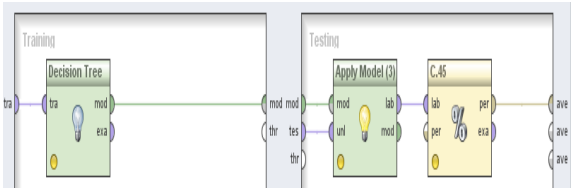
Sumber: (Sutoyo, 2018)

Gambar 3. Kerangka Kerja Proses Utama

Berikut ini keterangan dari kerangka kerja untuk pembentukan model.

1. Attribut-attribut dari dataset yang digunakan adalah JUMLAH_SESI dan NILAI_RATA-RATA.
2. Algoritma klasifikasi yang digunakan adalah *Decission Tree* C 4.5 dan *Random Forest*
3. Model validasi yang digunakan adalah *random 10-fold cross-validation*, artinya data training dibagi menjadi 10 bagian yang sama dan proses pembelajaran dilaksanakan sebanyak 10 kali.
4. Model evaluasi yang digunakan menggunakan akurasi sebagai indikator kinerja dari *classifier*.

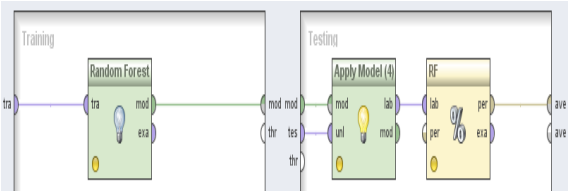
Operator C4.5 pada kerangka kerja proses utama memiliki sub proses seperti gambar berikut ini.



Sumber: (Sutoyo, 2018)

Gambar 4. Subproses C4.5

Operator *Random Forest* pada kerangka kerja proses utama memiliki sub proses seperti gambar berikut ini.



Sumber: (Sutoyo, 2018)

Gambar 5. Subproses *Random Forest*

Evaluation

Pada tahapan ini, model yang dihasilkan akan dievaluasi. Hasil evaluasi akan menentukan layak atau tidaknya model yang telah dihasilkan pada tahapan sebelumnya untuk digunakan.

Untuk evaluasi digunakan metode 10-*fold cross-validation*. Metode ini sangat handal untuk melaksanakan evaluasi terhadap model dikarenakan evaluasi dilaksanakan secara berulang-ulang. Meskipun tentu berkonsekuensi proses evaluasi memakan waktu lebih lama.

Deployment

Pada tahapan ini, model yang dinyatakan layak setelah melalui proses evaluasi akan digunakan untuk melaksanakan klasifikasi terhadap dataset. Pada penelitian akan dibuat aplikasi yang menerapkan *rule* yang terkandung pada model.

HASIL DAN PEMBAHASAN

Berdasarkan kerangka kerja yang telah dijelaskan pada bagian sebelumnya didapatkan hasil percobaan sebagai berikut:

Accuracy

Untuk mengukur akurasi digunakan *Confusion Matrix*. Berikut ini *Confusion Matrix* dari C4.5.

Tabel 1. *Confusion Matrix Classifier* C4.5

	true A	true B	true C	true D	precision (%)
pred A	29	1	0	0	96.67
pred B	0	29	2	0	93.55
pred C	0	0	28	1	96.55
pred D	0	0	0	33	100
recall (%)	100	96.67	93.33	97.06	

Sumber: (Sutoyo, 2018)

Penjelasan perhitungan akurasi dari *Confusion Matrix* C4.5 di atas adalah sebagai berikut:

pred A – true A: Jumlah record yang diprediksi masuk ke kelas A dan ternyata benar record tersebut termasuk kelas A sebanyak 29 record.

pred B – true B: Jumlah record yang diprediksi masuk ke kelas B dan ternyata benar record tersebut termasuk kelas B sebanyak 29 record.

pred C – true C: Jumlah record yang diprediksi masuk ke kelas C dan ternyata benar record tersebut termasuk kelas C sebanyak 28 record.

pred D – true D: Jumlah record yang diprediksi masuk ke kelas D dan ternyata benar

record tersebut termasuk kelas D sebanyak 33 record.

Jumlah dari A + B + C + D disebut *True Positive* (TP). 29 + 29 + 28 + 33 = 119. Jadi, jumlah record yang berhasil diprediksi dengan tepat oleh *Classifier* C4.5 sebanyak 119 record. Jumlah seluruh record adalah 123 record sehingga akurasi dihitung dengan (119/123)*100% = 96,73 %.

Untuk *Confusion Matrix* dari *Random Forest* disajikan pada tabel berikut ini.

Tabel 2. *Confusion Matrix Classifier* RF

	true A	true B	true C	true D	precision (%)
pred A	28	1	0	0	96.55
pred B	1	28	2	0	90.32
pred C	0	1	28	1	93.33
pred D	0	0	0	33	100
recall (%)	96.55	93.33	93.33	97.06	

Sumber: (Sutoyo, 2018)

Penjelasan perhitungan akurasi dari *Confusion Matrix Random Forest* di atas adalah sebagai berikut:

pred A – true A: Jumlah record yang diprediksi masuk ke kelas A dan ternyata benar record tersebut termasuk kelas A sebanyak 28 record.

pred B – true B: Jumlah record yang diprediksi masuk ke kelas B dan ternyata benar record tersebut termasuk kelas B sebanyak 28 record.

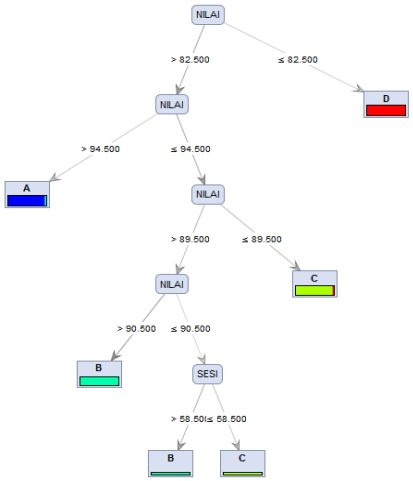
pred C – true C: Jumlah record yang diprediksi masuk ke kelas C dan ternyata benar record tersebut termasuk kelas C sebanyak 28 record.

pred D – true D: Jumlah record yang diprediksi masuk ke kelas D dan ternyata benar record tersebut termasuk kelas D sebanyak 33 record.

Jumlah dari A + B + C + D disebut *True Positive* (TP). 28 + 28 + 28 + 33 = 117. Jadi, jumlah record yang berhasil diprediksi dengan tepat oleh *Classifier Random Forest* sebanyak 117 record. Jumlah seluruh record adalah 123 record sehingga akurasi dihitung dengan (117/123)*100% = 95,13 %.

Knowledge

Pengetahuan yang dihasilkan oleh algoritma C4.5 dapat dipresentasikan dalam dua bentuk, yaitu pohon keputusan dan aturan menggunakan IF-THEN.



Sumber: (Sutoyo, 2018)

Gambar 6. Model Pohon C4.5

Pohon keputusan tersebut dibaca dari atas ke bawah atau dari akar (simpul pertama paling atas) sampai ke daun (simpul terluar yang tidak lagi memiliki cabang). Berikut cara membacanya dengan mengacu pada tiap simpulnya.

Jika NILAI 82.5 ke bawah maka lansung masuk KELAS D tanpa perlu lagi melihat jumlah SESI pengerjaan.

Jika NILAI di atas 82.5 maka ada dua kemungkinan. Pertama, jika NILAI di atas 94.5 maka lansung masuk KELAS A tanpa perlu lagi melihat jumlah SESI pengerjaan. Kedua, jika NILAI 94.5 ke bawah atau dengan kata lain NILAI di atas 82.5 namun 94.5 ke bawah, yakni 82.5 < NILAI ≤ 94.5 maka ada dua kemungkinan lagi.

Pertama, jika NILAI 89.5 ke bawah atau dengan kata lain NILAI di atas 82.5 namun 89.5 ke bawah, yakni 82.5 < NILAI ≤ 89.5 maka lansung masuk KELAS C tanpa perlu lagi melihat jumlah SESI pengerjaan. Kedua, jika NILAI di atas 89.5 atau dengan kata lain NILAI di atas 89.5 namun 94.5 ke bawah, yakni 89.5 < NILAI ≤ 94.5 maka ada dua kemungkinan lagi.

Pertama, jika NILAI di atas 90.5 atau dengan kata lain NILAI di atas 90.5 namun 94.5 ke bawah, yakni 90.5 < NILAI ≤ 94.5 maka lansung masuk KELAS B tanpa perlu lagi melihat jumlah SESI pengerjaan. Kedua, jika NILAI 90.5 ke bawah atau dengan kata lain NILAI di atas 89.5 namun 90.5 ke bawah, yakni 89.5 < NILAI ≤ 90.5 maka ada dua kemungkinan lagi berdasarkan jumlah SESI pengerjaan.

Pertama, jika jumlah SESI di atas 58.5 atau dengan kata lain NILAI di atas 89.5 namun 90.5 ke bawah dan jumlah SESI di atas 58.5, yakni 89.5 < NILAI ≤ 90.5 AND SESI > 58.5 maka masuk KELAS B. Kedua, jika jumlah SESI 58.5 ke bawah atau dengan kata lain NILAI di atas 89.5 namun 90.5 ke

bawah dan jumlah SESI 58.5 ke bawah, yakni 89.5 < NILAI ≤ 90.5 AND SESI ≤ 58.5 maka masuk KELAS C.

Adapun untuk aturan untuk setiap KELAS berikut populasi record yang memenuhi aturan tersebut dapat digambarkan seperti pada *Rule* dari Model Pohon C4.5 berikut.

NILAI > 82.500
| NILAI > 94.500: A {A=29, B=1, C=0, D=0}
| NILAI ≤ 94.500
| | NILAI > 89.500
| | | NILAI > 90.500: B {A=0, B=26, C=0, D=0}
| | | NILAI ≤ 90.500
| | | | SESI > 58.500: B {A=0,B=3, C=0,D=0}
| | | | SESI ≤ 58.500: C {A=0,B=0,C=2,D=0}
| | NILAI ≤ 89.500: C {A=0, B=0, C=28, D=1}
NILAI ≤ 82.500: D {A=0, B=0, C=0, D=33}

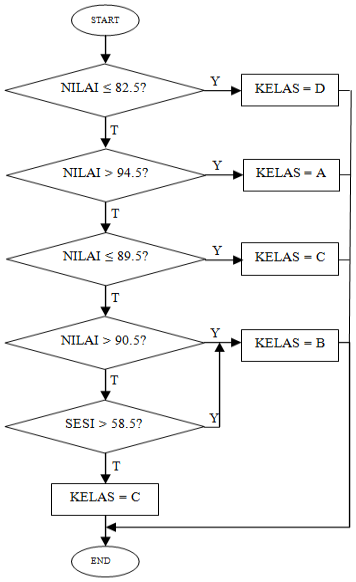
Aturan untuk KELAS A adalah NILAI > 94.5. Siswa akan masuk ke KELAS A jika NILAI di atas 94.5. Berdasarkan dataset, ada 29 record yang mengikuti aturan ini. Namun, ada 1 record yang menyimpang, yakni NILAI record tersebut di atas 94.5 tapi kenyataannya ia masuk ke KELAS B.

Aturan untuk KELAS B ada dua. Pertama, NILAI > 90.5. Siswa akan masuk ke KELAS B jika NILAI di atas 90.5. Berdasarkan dataset, ada 26 record yang mengikuti aturan ini. Kedua, NILAI ≤ 90.5 AND SESI > 58.5. Siswa akan masuk ke KELAS B meskipun NILAI 90.5 ke bawah dengan syarat jumlah SESI pengerjaannya di atas 58.5 kali. Berdasarkan dataset, ada 3 record yang mengikuti aturan ini.

Aturan untuk KELAS C ada dua. Pertama, 82.5 < NILAI ≤ 89.5. Siswa akan masuk ke KELAS C jika NILAI di atas 82.5 namun 89.5 ke bawah. Berdasarkan dataset, ada 28 record yang mengikuti aturan ini. Kedua, 89.5 < NILAI ≤ 90.5 AND SESI ≤ 58.5. Siswa akan masuk KELAS C jika jumlah SESI 58.5 ke bawah atau dengan kata lain NILAI di atas 89.5 namun 90.5 ke bawah dan jumlah SESI 58.5 ke bawah. Berdasarkan dataset, ada 2 record yang mengikuti aturan ini.

Aturan untuk KELAS D hanya ada 1, yaitu NILAI ≤ 82.5. Siswa akan masuk ke KELAS D jika NILAI 82.5 ke bawah. Berdasarkan dataset, ada 33 record yang mengikuti aturan ini.

Aturan tersebut dapat digambarkan menggunakan *flowchart* sebagai gambar 7 berikut.



Sumber: (Sutoyo, 2018)

Gambar 7. *Flowchart* dari Model Pohon C4.5

Alur logika dari *flowchart* yang merepresentasikan rule dari model sebagai berikut.

1. Uji apakah NILAI 82.5 kebawah. Jika dipenuhi maka masuk KELAS D.
2. Jika tidak memenuhi kondisi 1 maka uji lagi apakah NILAI di atas 94.5. Jika dipenuhi maka masuk KELAS A.
3. Jika tidak memenuhi kondisi 2 maka uji lagi apakah NILAI 89.5 kebawah. Jika dipenuhi maka masuk KELAS C.
4. Jika tidak memenuhi kondisi 3 maka uji lagi apakah NILAI di atas 90.5. Jika dipenuhi maka masuk KELAS B.
5. Jika tidak memenuhi kondisi 4 maka uji lagi apakah SESI di atas 58.5. Jika dipenuhi maka masuk KELAS B.
6. Jika tidak memenuhi kondisi 5 maka masuk KELAS C.

Pseudocode aturan dari *flowchart* tersebut adalah sebagai berikut.

```
if ($Nilai <= 82.5)
  {$Kelas = 'D';}
else if ($Nilai > 82.5)
  {
  if  ($Nilai > 94.5)
    {$Kelas = 'A';}
  else if ($Nilai <= 94.5)
    {
    if ($Nilai <= 89.5)
      {$Kelas = 'C';}
    else if ($Nilai > 89.5)
      {
      if ($Nilai > 90.5)
        {$Kelas = 'B';}
      else if ($Nilai <= 90.5)
        {
        if ($Sesi > 58.5)
          {$Kelas = 'B';}
```

```
        else if ($Sesi <= 58.5)
          {$Kelas = 'C';}
        }
      }
    }
  }
```

Implementasi

Untuk mengimplementasikan *rule* dapat dibuat aplikasi yang mampu melaksanakan prediksi dari dataset sesuai *rule*. Berikut contoh tampilan untuk masukan data.

Sumber: (Sutoyo, 2018)

Gambar 9. Interface Aplikasi Klasifikasi

Untuk Single Record Test, masukkan nilai, jumlah pengerjaan sesi, dan kelas kemudian klik Kirim. Untuk Multi Record Test, pilih file dalam format csv kemudian klik Upload CSV. Contoh hasilnya pada table 3 sebagai berikut.

Tabel 3. Contoh Hasil Klasifikasi

25	95	44	A	A	TRUE
26	95	41	A	A	TRUE
27	95	41	A	A	TRUE
28	95	40	A	A	TRUE
29	95	40	A	A	TRUE
30	95	29	B	A	FALSE
31	94	62	B	B	TRUE
32	94	53	B	B	TRUE
33	94	49	B	B	TRUE
34	94	45	B	B	TRUE
35	94	42	B	B	TRUE

Sumber: (Sutoyo, 2018)

Aplikasi akan menampilkan hasil pengolahan data yang diinputkan menggunakan rules yang telah dikodekan. Jika prediksi KELAS dari aplikasi sama dengan KELAS yang tercantum di dataset maka hasilnya TRUE. Jika hasil prediksi tidak sama maka hasilnya FALSE dan diberi highlight warna merah.

KESIMPULAN

Berdasarkan percobaan didapatkan bahwasanya algoritma *Decision Tree* yang diujicoba menunjukkan hasil yang memuaskan. Baik C4.5 maupun *Random Forest* telah menunjukkan kinerja yang tinggi dalam ukuran akurasi, yakni 97,63% untuk C4.5 dan 95,13% untuk *Random Forest*. Berdasarkan ukuran akurasi ini, C4.5 mengungguli *Random Forest* sebesar 2,5%. Oleh karena itu, model dan *rule* yang dihasilkan oleh C.45 digunakan sebagai dasar pengembangan prototipe aplikasi klasifikasi. Untuk penelitian selanjutnya, dapat digunakan dataset yang lebih besar dengan jumlah record yang lebih banyak. Selain itu juga penelitian dapat dikembangkan dengan penggunaan atribut yang lebih banyak.

REFERENSI

Ahmed, A. B. E. D., & Elaraby, I. S. (2014). Data Mining: A prediction for Student's Performance Using Classification Method. *World Journal of Computer Application and Technology*, *2*(2), 43–47. https://doi.org/10.13189/WJCAT.2014.020203

Cerón, J. D., López, D. M., & Eskofier, B. M. (2018). Human Activity Recognition Using Binary Sensors, BLE Beacons, an Intelligent Floor and Acceleration Data: A Machine Learning Approach. *Proceedings*, *2*(19), 1265. https://doi.org/10.3390/proceedings2191265

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining Concepts and Techniques* (3rd Editio). Waltham, USA: Morgan Kaufmann Publishers.

Larasati, D. A. H. D., & Sutrisno, T. (2018). Tourism Site Recommendation in Jakarta Using Decision Tree Method Based on Web Review, 195–209.

North, M. A. (2012). *Data Mining for the Masses*. *Computer Global Text Project*. Georgia: Global Text Project.

Saber Iraji, M., Aboutalebi, M., Seyedaghaee, N. R., & Tosinia, A. (2012). Students Classification With Adaptive Neuro Fuzzy. *International Journal of Modern Education and Computer Science*, *4*(7), 42–49. https://doi.org/10.5815/ijmecs.2012.07.06

Sutoyo, I. (2018). *Laporan Akhir Penelitian "Implementasi Algoritma Decision Tree Untuk Klasifikasi Data Peserta Didik."* Jakarta.

Wambui, E., George, N., & Kimani, S. (2018). An Intelligent Model for Fleet Management by Use of Sensor Enabled Tags Integrated With GPRS Technology, *4*(11), 30–40. https://doi.org/10.31695/IJASRE.2018.329

08

Yadav, S. K., & Pal, S. (2012). Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification. *World of Computer Science and Information Technology Journal (WCSIT)*, *2*(2), 51–56. Retrieved from http://arxiv.org/abs/1203.3832

Ye, N. (2014). *Data Mining Theories, Algorithms, and Examples*. 6000 Broken Sound Parkway NW: Taylor & Francis Group, LLC.