PENERAPAN ALGORITMA K-MEANS DAN C4.5 UNTUK CLUSTERING JURUSAN SISWA BARU PADA SEKOLAH MENENGAH KEJURUAN (STUDI KASUS: SMK NEGERI 1 PARON)

Disusun sebagai salah satu syarat memperoleh Gelar Strata I
pada Jurusan Informatika Fakultas Komunikasi dan Informatika

Oleh:
AZZAHRA SALSABILLA
L200170130

PROGRAM STUDI INFORMATIKA
FAKULTAS KOMUNIKASI DAN INFORMATIKA
UNIVERSITAS MUHAMMADIYAH SURAKARTA
2021

HALAMAN PERSETUJUAN

PENERAPAN ALGORITMA K-MEANS DAN C4.5 UNTUK CLUSTERING JURUSAN SISWA BARU PADA SEKOLAH MENENGAH KEJURUAN (STUDI KASUS: SMK NEGERI 1 PARON)

PUBLIKASI ILMIAH

Oleh:
AZZAHRA SALSABILLA
L200170130

Telah diperiksa dan disetujui untuk diuji oleh:

Dosen Pembimbing

Azizah Fatmawati, S.T., M.Cs

NIK.1198

HALAMAN PENGESAHAN

PENERAPAN ALGORITMA K-MEANS DAN C4.5 UNTUK CLUSTERING JURUSAN SISWA BARU PADA SEKOLAH MENENGAH KEJURUAN (STUDI KASUS: SMK NEGERI 1 PARON)

OLEH

AZZAHRA SALSABILLA

L200170130

Telah dipertahankan di depan Dewan Penguji
Fakultas Komunikasi dan Informatika
Universitas Muhammadiyah Surakarta
Pada hari Senin, 19 Juli 2021
dan dinyatakan telah memenuhi syarat

Dewan Penguji:

1. **Azizah Fatmawati, S.T., M.Cs.** (……..)
   (Ketua Dewan Penguji)
2. **Devi Afriyantari Puspa Putri, S.Kom., M.Sc.** (……..,
   (Anggota 1 Dewan Penguji)
3. **Maryam, S.Kom., M.Eng.** (……..)
   (Anggota 2 Dewan Penguji)

Dekan
Fakultas Komunikasi dan Informatika

Nurgiyatna, S.T., M.Sc., Ph.D.
NIK. 881

PERNYATAAN

Dengan ini saya menyatakan bahwa dalam publikasi ilmiah ini tidak terdapat karya yang pernah diajukan untuk memperoleh gelar kesarjanaan di suatu perguruan tinggi dan sepanjang pengetahuan saya juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan orang lain, kecuali secara tertulis diacu dalam naskah dan disebutkan dalam daftar pustaka.

Apabila kelak terbukti ada ketidakbenaran dalam pernyataan saya di atas, maka akan saya pertanggungjawabkan sepenuhnya.

Surakarta, 19 Juli 2021
Penulis

AZZAHRA SALSABILLA
L200170130

PENERAPAN ALGORITMA K-MEANS DAN C4.5 UNTUK CLUSTERING JURUSAN SISWA BARU PADA SEKOLAH MENENGAH KEJURUAN (STUDI KASUS: SMK NEGERI 1 PARON)

Abstrak

Jurusan di Sekolah Menengah Kejuruan (SMK) digunakan untuk rujukan dalam menyalurkan kemampuan siswa. Penjurusan dilakukan saat proses pendaftaran siswa baru. Pada umumnya calon siswa harus memilih jurusan sesuai dengan rata-rata nilai rapor atau nilai ujian nasional Sekolah Menengah Pertama (SMP) tanpa melihat kriteria lain sehingga mengakibatkan terjadinya kesalahan pemilihan jurusan yang tidak sesuai dengan minat dan kemampuannya. Penelitian ini bertujuan membentuk klasifikasi jurusan siswa secara efektif dan efisen. Metode dalam penelitian ini yaitu teknik Data Mining yang menerapkan algoritma K-Means dan algoritma C4.5. Algoritma K-Means digunakan dalam proses pengelompokan jurusan berdasarkan kriteria pendukung berupa nilai rapor SMP yaitu Nilai Matematika dan Nilai Bahasa, rata-rata penghasilan orang tua, serta jumlah saudara calon siswa. Hasil dari algoritma K-Means akan dibandingkan dengan hasil dari algoritma C4.5 yang berbentuk rule (pohon keputusan). Penelitian ini menghasilkan nilai akurasi dari algoritma K-Means sebesar 58,22%, sedangkan algoritma C4.5 menghasilkan nilai Precision sebesar 26,44%, Recall sebesar 25,9%, dan Accuracy sebesar 41,61%,

Kata Kunci: Algoritma C4.5, Algorima K-Means, Klasifikasi, Kriteria, Penjurusan

Abstract

Courses in Vocational High Schools (SMK) are used as a reference to channel the abilities of students. The cornering is done during the registration of new students. In general, prospective students must choose a major by the average score of the report card or national test scores of Junior High School (SMP) regardless of other criteria resulting in the occurrence of mistakes in the selection of majors that are not by their interests and abilities. This research aims to classify courses for students appropriately, effectively, and efficiently. The method used in this study is with Data Mining which applies K-Means Algorithm and C4.5 Algorithm. K-Means algorithm is used to group courses based on supporting criteria in the form of junior high school grades, namely Math and Language Scores, average parental income, and several prospective students' siblings. The results of the K-Means algorithm will be compared to the results of the C4.5 algorithm in the form of a rule (decision tree). This study produced an accuracy value from K-Means algorithm by 58,22%, while the C4.5 algorithm resulted in an Accuracy value of 41,61%, Precision of 26,44%, and Recall value of 25,9%.

Keywords: C4.5 Algorithm, K-Means Algorithm, Classification, Criteria, Cornering

1. PENDAHULUAN

Pendidikan merupakan kebutuhan utama dalam mengembangkan potensi diri. Perkembangan teknologi yang semakin maju membuat pendidikan menjadi bekal yang sangat penting untuk mengantarkan kita menuju masa depan yang cerah (Umar et al., 2017). Pendidikan kejuruan adalah sistem pendidikan menengah yang mempersiapkan siswa dan siswinya agar memiliki kemampuan yang lebih untuk bekerja pada suatu bidang pekerjaan tertentu. Pendidikan yang berbasis kompetensi menekankan sistem pembelajaran pada kemampuan yang dimiliki siswa sehingga dapat mengikuti kompetensi dan kurikulum sesuai jurusan (Firza & Sarjono, 2020).

SMK Negeri 1 Paron merupakan sekolah kejuruan yang bertempat di Jalan Raya Gentong Kecamatan Paron Kabupaten Ngawi. Terdapat beberapa jurusan di SMK Negeri 1 Paron yaitu Akuntansi, Pemasaran, Perhotelan, Teknik Kendaraan Ringan (TKR), dan Teknik Sepeda Motor (TSM). Proses penjurusan di sekolah tersebut masih terbilang kurang tepat, karena siswa dapat masuk pada jurusan hanya berdasarkan satu kriteria yaitu nilai Ujian Nasional SMP. Sehingga menyebabkan siswa tidak masuk pada jurusan yang sesuai minat dan bakatnya. Perlu adanya suatu klasifikasi berdasarkan kriteria lain untuk mengatasi permasalahan tersebut. Cara klasifikasi yang digunakan yaitu teknik data mining dengan mengimplementasikan algoritma K-Means dan C4.5.

Menurut (Bustami, 2010), data mining adalah metode menggunakan sejumlah data untuk mendapatkan informasi berharga yang sebelumnya tidak diketahui dan menggunakannya untuk pengambilan keputusan penting. Data mining juga dikenal sebagai penemuan pengetahuan (KDD) atau pengenalan pola dalam *database*. Istilah KDD mengacu pada penemuan pengetahuan data karena tujuan utama dari data mining adalah untuk memproses data dalam *database* dan menghasilkan informasi baru yang berguna. Istilah pengenalan pola ditujukan untuk mengekstraksi pengetahuan dalam potongan data yang akan dihadapi (Nur et al., 2017). Adapun proses KDD yaitu *Data Selection* (pemilihan data), *Preprocessing* (proses *cleaning* sehingga data atau informasi yang digunakan relevan), *Transformation* (proses *coding* sehingga data yang akan digunakan sesuai dengan proses dalam data mining), *Data Mining* (proses mencari pola menggunakan metode tertentu), dan terakhir yaitu *Interpretation/Evaluation* (proses

pemeriksaan kesesuaian pola dengan fakta dan hipotesa) (Kurniasari & Fatmawati, 2019).

Clustering adalah proses pengelompokan data ke dalam kelompok-kelompok dengan kesamaan data dalam kelompok tinggi dan kesamaan data antar kelompok rendah (Dharmayanti et al., 2017; Mardiani, 2015; Tan et al., 2016). K-Means adalah metode pengelompokan data yang mempartisi data menjadi satu atau lebih kelompok. Algoritma K-Means memiliki keunggulan dalam eksekusi yang relatif cepat, implementasi yang sederhana, dan penggunaan yang luas dalam tugas data mining (Tandy & Assegaff, 2019). Algoritma K-Means adalah algoritma pengelompokan data berulang yang membagi data menjadi sejumlah K *cluster* yang telah ditentukan (Kantardzic, 2011).

Algoritma C4.5 digunakan untuk mengubah data menjadi pohon keputusan yang mewakili suatu aturan. Algoritma C4.5 dapat menghasilkan pohon keputusan yang memiliki tingkat akurasi yang tinggi dan memiliki keunggulan yaitu mudah diinterpretasikan untuk pemecahan masalah serta mampu menghitung sifat-sifat diskrit dan numerik (Wahyuni, 2018). Tujuan dari pembuatan pohon keputusan adalah untuk mempermudah penyelesaian masalah. Konsep pohon keputusan adalah mengubah data menjadi pohon keputusan dan aturan keputusan (Faisal, 2019).

Beberapa penulis telah menerapkan teknik *clustering* menggunakan algoritma K-Means dan C4.5 dalam hal pengelompokkan data seperti (Nasari & Darma, 2015) dalam penelitiannya menuliskan bahwa tujuan penelitian yaitu menerapkan metode pengelompokkan K-Means pada data siswa baru tahun ajaran 2014/2015. Hasil penelitian di dapatkan bahwa asal sekolah SMA rata-rata mengambil jurusan sistem informasi, sedangkan yang berasal dari sekolah SMK mengambil jurusan Informatika. Metode di atas cukup efektif dan efisien dalam proses pengelompokkan data.

Dalam penelitian pengukuran luas wilayah hutan mangrove yang membandingkan metode segmentasi K-Means dan *region growing* menunjukkan hasil bahwa metode sementasi menggunakan k-means mendapatkan nilai akurasi yang lebih baik dibanding akurasi dari *region growing*. K-Means *cluster*ing mendapatkan nilai akurasi sebesar 59,26% dan nilai *region growing* 33,33%. Metode sementasi tersebut dapat digunakan dengan baik untuk menghitung luas wilayah sehingga mendapatkan hasil yang akurat (Cerah et al., 2019).

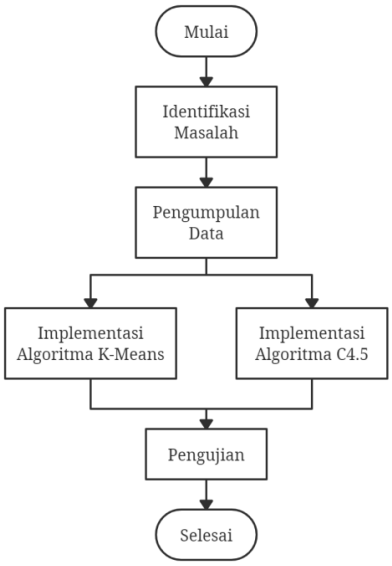
Menurut penelitian yang dilakukan oleh (Ong, 2013), algoritma K-Means dapat membantu pihak marketing universitas dalam proses pencarian hingga mendapatkan calon mahasiswa baru dari berbagai daerah serta membantu dalam proses pemasaran. Dalam penelitian yang lain yang membandingkan berbagai algoritma untuk membentuk pohon keputusan dan mendapatkan kesimpulan algoritma C4.5 adalah algoritma dengan performa terbaik dan nilai akurasi tertinggi (Mohankumar et al., 2016).

Menurut (Jadhav & Channe, 2016) perbandingan performa dari pohon keputusan, Naïve Bayes, dan K-NN didapatkan hasil bahwa pohon keputusanlah yang memiliki nilai *error rate* yang rendah dan performa paling cepat dibanding dengan metode lainnya. Dalam penelitian lain yang menggunakan atribut seperti IPK, lama studi, asal daerah, nilai TOEFL mendapatkan hasil nilai akurasi 60,52%, *recall* 60,73%, dan presisi sebesar 63.93%. Sehingga disimpulkan bahwa algoritma yang baik untuk melakukan prediksi terhadap kelulusan mahasiswa yaitu algoritma C4.5 (Putri & Waspada, 2018).

Dari beberapa penelitian sebelumnya didapatkan kesimpulan bahwa algoritma K-Means adalah teknik pengelompokan data tanpa pemantauan dan teknik data mining yang umum dan mudah digunakan untuk proses pengelompokkan yang efektif dan efisien (Nasari & Darma, 2015). Sedangkan dari penelitian mengenai algoritma C4.5 dapat disimpulkan bahwa algoritma tersebut memiliki nilai akurasi yang relatif tinggi dan performa yang cepat. Hal tersebut yang menjadi dasar penulis melakukan kegiatan penelitian dengan menerapkan algoritma K-Means dan C4.5 dalam proses pengelompokan jurusan bagi siswa SMKN 1 Paron.

2. METODE

Metode penelitian atau langkah penelitian yang diterapkan dalam penelitian ini terdapat beberapa proses yaitu proses yang pertama identifikasi masalah, selanjutnya proses pengumpulan data, implementasi algoritma, dan yang terakhir yaitu proses pengujian dari kedua algoritma yang digunakan dalam proses klasifikasi jurusan. Implementasi algoritma dilakukan menggunakan 2 algoritma yaitu algoritma K-Means dan C4.5 yang keduanya akan dijalankan menggunakan *software* RapidMiner Studio. Adapun gambaran dari langkah-langkah penelitian tersebut ditunjukkan oleh Gambar 1.



Gambar 1. Langkah-langkah Penelitian

2.1 Identifikasi Masalah

Tahapan pertama dalam penelitian ini adalah identifikasi masalah yang dilakukan melalui proses wawancara dengan pihak sekolah baik guru maupun siswa. Terdapat permasalahan dalam proses penjurusan yaitu kurang sesuainya minat dan jurusan yang didapatkan oleh siswa. Hal tersebut disebabkan karena kriteria yang digunakan dalam proses penjurusan yang kurang.

2.2 Pengumpulan Data

Pengumpulan data untuk penelitian ini dilakukan sesuai dengan dokumen siswa baru SMK Negeri 1 Paron. Data yang digunakan adalah data tahun ajaran 2020 hingga 2021. Data yang diperoleh sebanyak 265 record data yang terbagi menjadi 2 data yaitu data *training* 70% dari data *testing* 30%. Ada juga data dari tinjauan pustaka jurnal dan buku yang terkait dengan topik penelitian ini.

Seluruh data yang terkumpul masuk ke dalam data pra-olahan (*preprocessing data*). Pemilihan data pada tahap *preprocessing* didasarkan pada kriteria yang telah ditentukan sebelumnya. Beberapa faktor yang mempengaruhi kualitas dari suatu data, antara lain akurasi, integritas, konsistensi, realitas dan interpretasi (Dharmayanti et al., 2017).

2.3 Model yang diusulkan

Proses penentuan jurusan siswa SMK berdasarkan kriteria pendukung dilakukan dengan algoritma K-Means dan C4.5. Algoritma K-Means memiliki sekelompok langkah meliputi (Nur et al., 2017) :

a. Menentukan nilai centroid dan jarak yang akan digunakan. Menentukan nilai centroid menggunakan Persamaan (1).

$$\frac{Jumlahdata}{Jumlahclass+1} \quad (1)$$

b. Gunakan metrik jarak yang telah ditentukan untuk memetakan semua data ke centroid terdekat.
c. Menghitung kembali nilai centroid baru berdasarkan data tiap *cluster*.
d. Mengulang tahap 2 dan 3 sehingga mencapai nilai yang konvergen.

Menghitung nilai *Euclidean Distance* menggunakan Persamaan (2).

$$D(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (2)$$

Secara umum algoritma C4.5 dalam membangun suatu pohon keputusan yaitu memilih variabel yang akan digunakan sebagai *root*. Selanjutnya membuat cabang dari *root* untuk masing-masing nilai. Langkah terakhir yaitu membagi variabel terpilih di dalam cabang. Ulangi langkah-langkah tersebut sehingga variabel terpilih pada cabang terdapat dalam satu kelas (Oscario et al., 2019). Variabel yang digunakan sebagai akar dilihat berdasarkan nilai *gain* yang tertinggi dari variabel yang ada. Untuk menghitung nilai gain menggunakan Persamaan (3) (Prasatya et al., 2020).

$$Gain(S,A) = Entropy(S) - \sum_{i=1}^{n}\frac{|S_i|}{|S|} * Entropy(S_i) \quad (3)$$

Rumus di atas digunakan untuk menghitung nilai *gain*, sehingga didapatkan nilai gain yang tertinggi dari tiap variabel. Sementara itu, diperlukan nilai *entropy* dalam proses C4.5 ini. Untuk menghitung nilai *entropy* menggunakan Persamaan (4).

$$Entropy(S) = \sum_{i=1}^{n} -pi * log\,2\,pi \quad (4)$$

Kemudian dilakukan pengulangan dalam perhitungan *gain* dan *entropy* sampai seluruh record data terpartisi. Proses dari pemecahan pohon keputusan akan selesai apabila seluruh data telah mendapatkan kelas yang sama, di mana tidak ada variabel yang dapat dibagi kembali dan tidak ada data di dalam cabang yang belum terisi (Oscario et al., 2019).

2.4 Implementasi Algoritma

Tahap implementasi algoritma K-Means mengelompokkan data berdasarkan model yang diusulkan di atas. Hasil perhitungan K-Means akan dibandingkan dengan hasil dari algoritma C4.5. Perhitungan algoritma C4.5 akan menghasilkan pohon keputusan. Implementasi K-Means dan C4.5 dilakukan dengan menggunakan *software* RapidMiner Studio. Setiap perhitungan siswa akan membentuk klaster/kelompok besar.

2.5 Pengujian

Tahapan terakhir dalam penelitian ini yaitu pengujian. Dari hasil algoritma K-Means dan C4.5 akan dihitung nilai akurasi dari setiap algoritma tersebut. Hasil *cluster* dari algoritma K-Means akan dihitung nilai akurasi berdasarkan prediksi benar dan prediksi salah pada tiap iterasi. Sedangkan hasil pohon keputusan dari C4.5 akan uji dengan mencari nilai *Accuracy*, *Precision*, dan *Recall* dari tiap *role*. Nilai akurasi dari kedua algoritma akan dibandingkan sehingga ditemukan pengujian yang terbaik.

3. HASIL DAN PEMBAHASAN

3.1 Proses Algoritma K-Means

Dalam perhitungan K-Means menggunakan data *testing* dari set data yang telah disiapkan. Proses ini akan menentukan sebanyak 5 kelompok (*cluster*). Ketentuan dari jumlah kelompok tersebut sesuai dengan kebutuhan yaitu jumlah jurusan yang ada di sekolah. Perhitungan dilakukan menggunakan *software* RapidMiner Studio. Perhitungan ini dilakukan untuk mendapatkan kelompok yang akan membantu proses penjurusan bagi siswa baru. Adapun beberapa atribut atau variabel yang digunakan dalam proses penjurusan yaitu :

a. Minat siswa dengan variabel Akuntansi, Pemasaran, Perhotelan, TKR, dan TSM
b. Nilai rapor siswa yaitu nilai Matematika dan Nilai Bahasa Indonesia
c. Penghasilan orang tua siswa, dilihat berdasarkan kebutuhan siswa ditiap jurusan yang berbeda seperti dana untuk praktikum jurusan
d. Jumlah saudara siswa

Adapun data dalam penelitian ini terbagi menjadi 2 yaitu terdapat data *testing* dan data *training*. Di mana jumlah data *testing* sebesar 30% dan data *training* sebesar 70%. Data *training* akan ditunjukkan pada Tabel 1, sedangkan data *testing* terdapat pada Tabel 2.

Tabel 1. Data Training

No	Nama	Penghasilan Ortu	Rombel Saat Ini	Jumlah Saudara	MINAT	MTK	INDO
1	Fatur April Fahrozi	1,000,000	BDP A	2	TKRO	87	92
2	Febri Aji Saputra	800,000	TBSM B	1	TBSM	87	87
3	Febri Cahyono	1,000,000	TKRO B	1	TKRO	85	85
4	Febri Khoirut Tamami	1,000,000	TBSM B	2	AKL	85	96
5	Febri Wulan Ningrum	1,000,000	AKL B	2	AKL	89	91
6	Febrian Dwi Ragil Setiyo	6,000,000	TBSM B	2	AKL	86	89
7	Feri Afandi	500,000	PH A	2	TKRO	80	91
8	Fery Dian Saputra	1,000,000	TKRO B	6	TKRO	88	92
9	Fiki Rahma Afandi	2,000,000	TBSM B	2	TBSM	93	90
10	Fiky Septian	1,000,000	TKRO B	1	TKRO	90	88
….	……	……..	……..	……	……	……..	……..
182	Yuliani Diana Safitri	1,000,000	BDP C	2	TKRO	80	85
183	Yunda Eka Apriliani	700,000	AKL D	1	TKRO	90	87
184	Zahrul Saphittro	1,000,000	TBSM C	2	TBSM	88	88
185	Zainal Abidin	1,000,000	TKRO C	1	TKRO	90	96
186	Zainul Anam	1,000,000	TKRO C	1	TBSM	92	90

Tabel 2. Data Testing

No	Nama	Penghasilan Ortu	Rombel Saat Ini	Jumlah Saudara	MINAT	MTK	INDO
1	A'an Afrizal Kurniawan	1,000,000	TKRO A	1	AKL	90	91
2	Achmad Rizki Fatoni	2,000,000	TKRO A	2	TKRO	85	80
3	Adellia Intan Juniarti	800,000	AKL A	1	AKL	90	80
4	Adistiya Andriani	1,500,000	BDP A	2	AKL	76	80
5	Adit Wahana Nur Rohim	4,000,000	BDP A	1	AKL	80	93
6	Adzain Wahyu Yogi Pratama	1,500,000	TKRO A	1	TKRO	88	90
7	Agus Rahmat Adhi Pratama	1,800,000	TKRO A	1	TKRO	87	76
8	Ahmad Arif Ali Mujahidin	2,000,000	TKRO A	2	TKRO	90	90
9	Ahmad Munawir Al Masduki	1,500,000	TKRO A	2	AKL	85	78
10	Ajeng Amelia Putri	1,000,000	BDP A	1	TKRO	79	85
….	……	……..	……..	……..	……	……..	……..
77	Erlangga Ananta Sudarto	1,000,000	TBSM A	1	AKL	89	88
78	Erni Tri Rahmawati	5,000,000	BDP A	3	BDP	88	89
79	Fadhil Wahyu Saputra	3,500,000	TBSM B	1	TKRO	90	89

Data *testing* akan dinormalisasi dengan skala dari variabel Jumlah Saudara. Proses normalisasi dilakukan agar seluruh variabel memiliki skala yang sama. Proses normalisasi dilakukan menggunakan rumus *Min-Max Normalization* ditunjukkan dalam Persamaan 5 dan hasil dari normalisasi terdapat pada Tabel 3.

$$v^1 = \frac{(v - min_a)}{max_a - min_a}(newmax_a - newmin_a) + newmin_a \quad (5)$$

Keterangan:

v^1 = Data ternormalisasi (data baru)

v = Data awal

min_a = nilai minimum skala awal

max_a = nilai maksimum skala awal

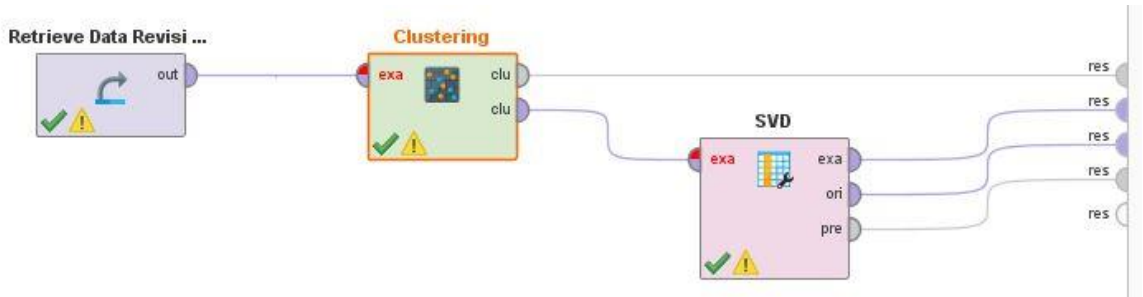
$newmin/max_a$ = nilai minimum/maksimum skala baru

Tabel 3. Hasil Normalisasi Data Testing

No	Penghasilan Ortu	Jumlah Saudara	Nilai	
			MTK	INDO
1	1.3	1	3.1	3.3
2	1.7	2	2.4	1.6
3	1.2	1	3.1	1.6
4	1.5	2	1	1.6
5	2.5	1	1.6	3.6
6	1.5	1	2.8	3.1
7	1.6	1	2.7	1
8	1.7	2	3.1	3.1
9	1.3	1	3.1	3.3
10	1.7	2	2.4	1.6
11	1.5	2	3.4	2.8
12	1.1	2	3	3
13	1.1	2	3.3	3
….	……	……..	……	……
77	1.3	1	3	2.8
78	2.8	3	2.8	3
79	2.3	1	3.1	3

3.1.1 Desain Proses Algoritma K-Means

Algoritma K-Means akan dihitung pada *software* RapidMiner Studio. Pada proses ini pertama yaitu menentukan atribut yang akan digunakan sebagai *id*. Dalam penelitian ini atribut yang berfungsi sebagai *id* yaitu nama siswa. Desain dari proses algoritma K-Means ditunjukkan pada Gambar 2.



Gambar 2. Desain Algoritma K-Means pada RapidMiner Studio

Dalam desain algoritma K-Means pada RapidMiner Studio tersebut menggunakan operator *Clustering* dan *Singular Value Decomposition* (SVD). Operator *clustering* berfungsi menentukan kelompok yang akan dibentuk pada data yang dimasukkan pada *software*. Pada operator *clustering* menggunakan parameter k sebesar

5, yang berarti akan terbentuk 5 kelompok pada proses *clustering* tersebut. Sedangkan SVD berfungsi untuk menentukan *invers* selanjutnya yang dapat menyelesaikan sistem persamaan linier yang sedang berlangsung.

3.1.2 Hasil dari Algoritma K-Means

Berikut adalah hasil dari proses *clustering* menggunakan algoritma K-Means.

a. *Singular Value Decomposition* (SVD)

1) Nilai Eigenvalue

Nilai *eigenvalue* adalah suatu nilai yang menunjukkan pengaruh seberapa besar suatu variabel terhadap pembentukan atau pengelompokkan karakteristik sebuah *vector* atau matriks. Nilai *eigenvalue* ditunjukkan pada Gambar 3.

Component	Singular Value	Proportion of Singular Values	Cumulative Singular Values	Cumulative Proportion of Singular Va...
SVD 1	16607377.901	1.000	16607377.901	1.000
SVD 2	694.021	0.000	16608071.921	1.000
SVD 3	34.558	0.000	16608106.480	1.000
SVD 4	7.348	0.000	16608113.828	1.000

Gambar 3. Nilai Eigenvalue dari SVD

2) Nilai SVD *Vectors*

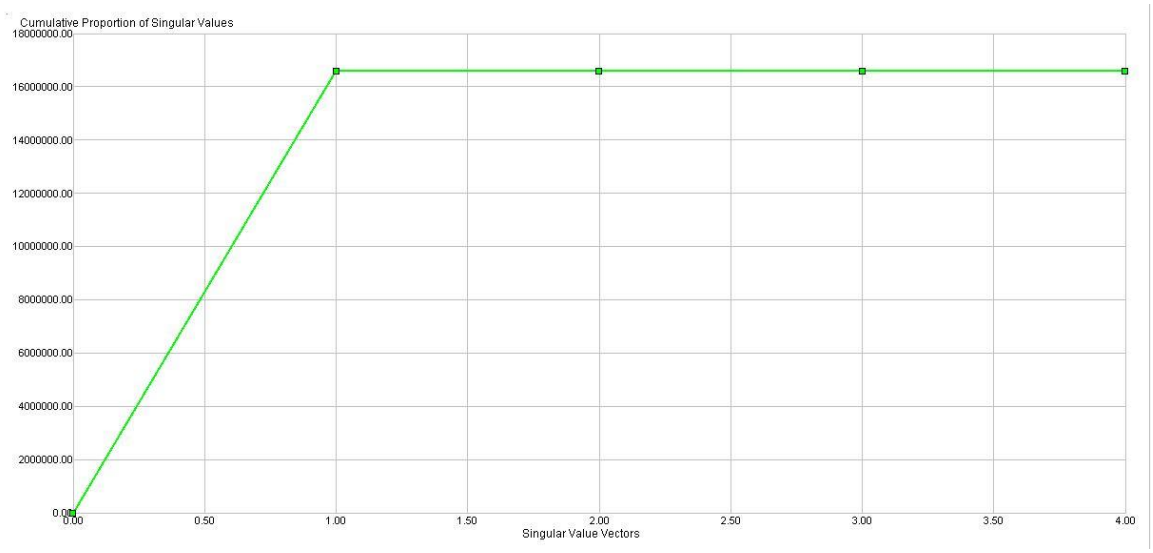
Nilai SVD *Vectors* berisi hasil-hasil SVD *vectors* dari setiap variabel yang dimasukkan ke dalam perhitungan. Nilai SVD *Vectors* ditunjukkan pada Gambar 4.

Attribute	SVD Vector 1	SVD Vector 2	SVD Vector 3
Penghasilan Ortu	1.000	-0.000	0.000
Jumlah Saudara	0.000	0.015	0.040
MTK	0.000	0.710	0.704
INDO	0.000	0.705	-0.709

Gambar 4. Nilai SVD Vectors

3) Nilai *Cumulative Variance*

Nilai dari *cumulative variance* di dalam *Singular Value Decomposition* (SVD) akan ditunjukkan pada Gambar 5.



Gambar 5. Nilai *cumulative variance*

b. *Example Set* (K-Means)

Hasil ini ditampilkan dalam mode *plot view* menggunakan diagram sebar untuk menentukan kelompok siswa (*cluster*) berdasarkan variabel input. Hasil *example set* (K-Means) ditunjukkan pada Gambar 6.

Gambar 6. Hasil *example set* (K-Means)

Hasil menunjukkan bahwa *cluster*_4 adalah kelompok siswa dengan nilai matematika lebih tinggi dari *cluster*_2 dan *cluster*_3. Oleh karena itu, hasil ini membantu dalam klasifikasi jurusan pada proses pengujian algoritma.

c. *Example Set* (SVD)

Tabel hasil *example set* (SVD) memungkinkan untuk menentukan pemisahan kelompok atau *cluster* siswa. Kolom Nama menampilkan nama siswa yang disertakan dalam data asli. Hasil *example set* (SVD) ditunjukkan pada Gambar 7.

Row No.	Nama	cluster ↑	svd_1
2	ACHMAD RIZ...	cluster_0	0.120
4	ADISTIYA AN...	cluster_0	0.090
6	ADZAIN WAH...	cluster_0	0.090
7	AGUS RAHM...	cluster_0	0.108
8	Ahmad Arif Ali...	cluster_0	0.120
9	Ahmad Muna...	cluster_0	0.090
11	AJENG KHAR...	cluster_0	0.090
14	Alfian Cindhi ...	cluster_0	0.090
17	ALIF MUHAZIZ	cluster_0	0.090
22	APRILRIYA R...	cluster_0	0.090
25	ARDIANFA C...	cluster_0	0.078
26	ARDIT CAHY...	cluster_0	0.120
32	ASTRI YUAN...	cluster_0	0.120
33	AULIA DAMAY...	cluster_0	0.090
37	BAGAS TRI S...	cluster_0	0.090

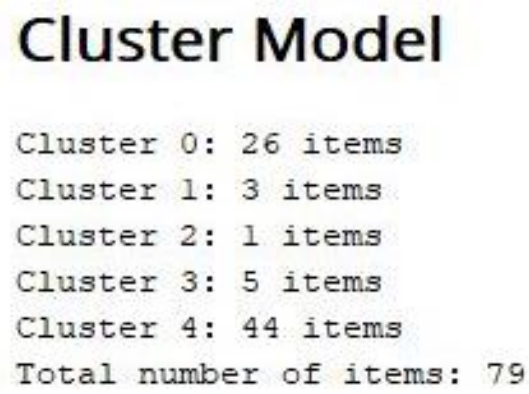
ExampleSet (79 examples, 2 special attributes, 1 regular attribute)

Gambar 7. Hasil *example set* (SVD)

d. *Cluster Model* (*Clustering*)

1) *Description*

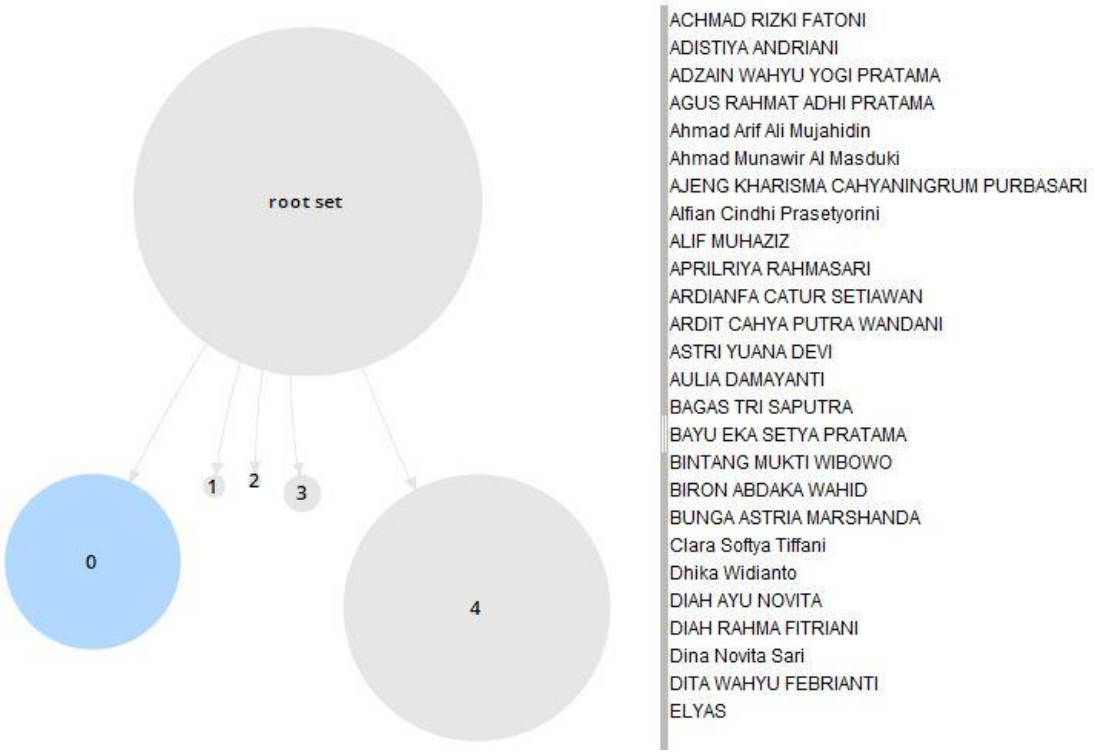
Pada model *cluster* dapat diketahui jumlah input data pada setiap *cluster*. *Cluster* 0 sebesar 26 orang, *Cluster* 1 sebesar 3 orang, *Cluster* 2 sebesar 1, dan *Cluster* 4 sebesar 44 orang. Dengan total data sebesar 79 siswa. Hasil model *cluster* berupa deskripsi ditunjukkan pada Gambar 8.



Gambar 8. Hasil *description cluster model*

2) *Graph*

Mode ini menunjukkan bentuk pembagian ke dalam kelompok-kelompok dengan pola pohon dan jenisnya. Setiap kelompok yang dikonfigurasi dapat dipilih dan ditampilkan anggota kelompok tersebut. Hasil untuk *cluster* 0 ditunjukkan pada Gambar 9.



Gambar 9. Hasil *graph* dari *cluster* 0

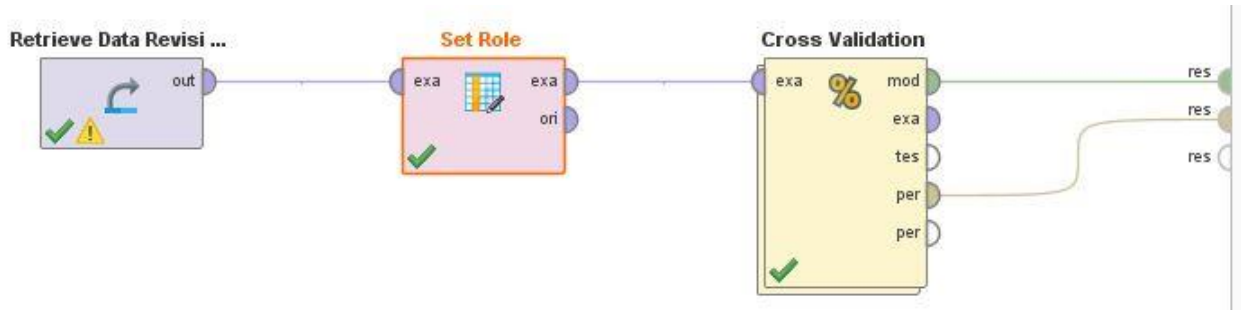
3.2 Proses Algoritma C4.5

Dalam perhitungan algoritma C4.5 menggunakan sebanyak data testing yang sama dengan sebelumnya. Perhitungan C4.5 ini sama dengan K-Means yaitu menentukan kelompok (*cluster*) yang berjumlah 5. Penentuan jumlah kelompok tersebut sesuai dengan kebutuhan yaitu jumlah jurusan yang ada di sekolah. Perhitungan Algoritma

C4.5 akan dilakukan menggunakan aplikasi RapidMiner. Atribut yang digunakan dalam penentuan jurusan sama dengan proses pada algoritma sebelumnya.

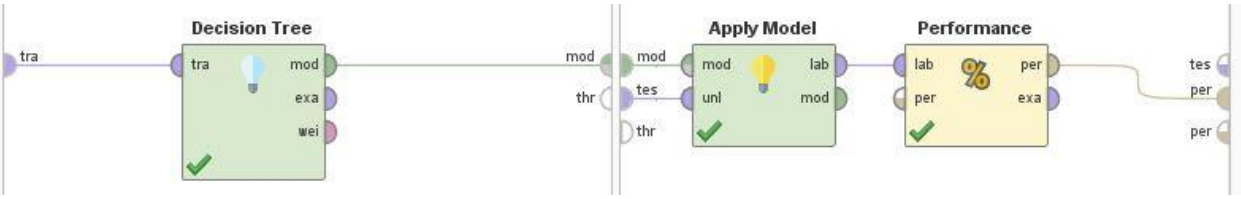
3.2.1 Desain Proses Algoritma C4.5

Algoritma C4.5 dihitung dengan *software* RapidMiner Studio. Proses pertama yaitu menentukan atribut yang akan diprediksi. Pada penelitian ini atribut tersebut yaitu jurusan yang akan diambil siswa apakah sesuai dengan minatnya. Desain dari proses C4.5 ditunjukkan pada Gambar 10.



Gambar 10. Desain C4.5 pada RapidMiner

Dari desain C4.5 pada Gambar 10, terdapat desain baru di dalam *Cross Validation*. Desain bagian dari *Cross Validation* ditunjukkan pada Gambar 11.

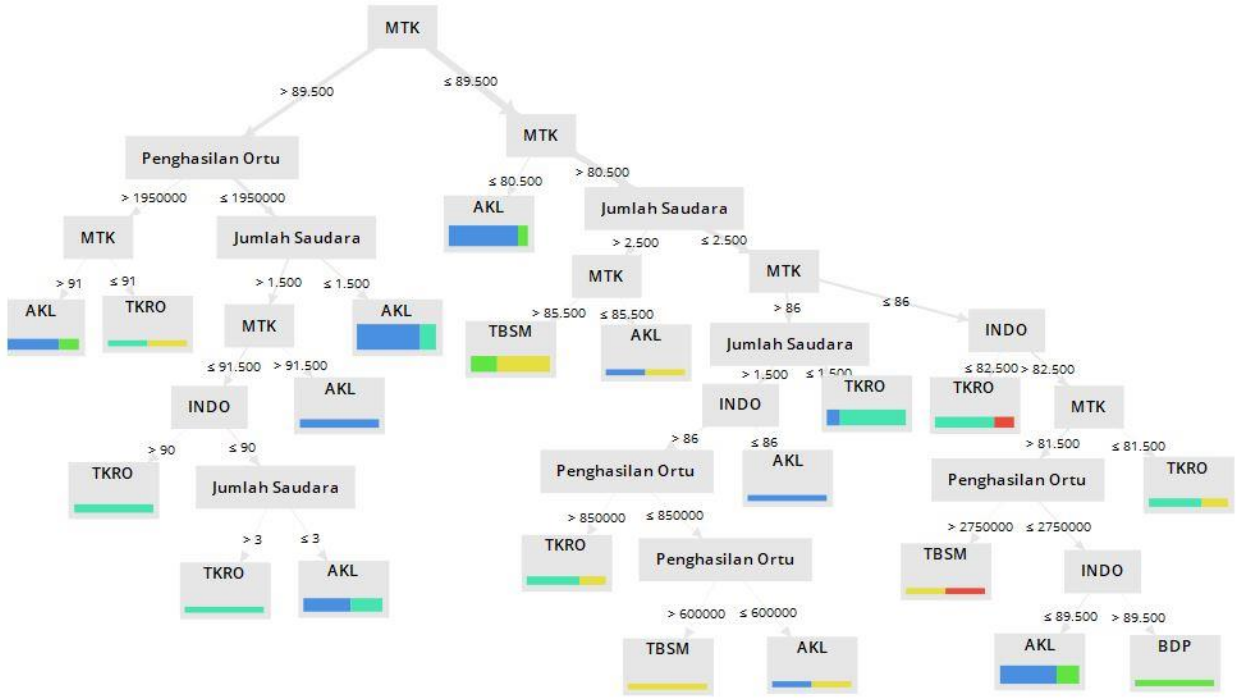


Gambar 11. Desain Bagian Cross Validation

Dalam desain Proses Algoritma C4.5 ini menggunakan operator antara lain *Set Role* dan *Cross Validation*. Di dalam *Cross Validation* terdapat operator yaitu *Decision Tree, Apply Model*, dan *Performance*.

3.2.2 Hasil dari Proses C4.5

Hasil yang didapatkan dari proses Algoritma C4.5 pada *software* RapidMiner Studio tinjukkan pada Gambar 12.



Gambar 12. Hasil RapidMiner Algoritma C4.5

Hasil pada Gambar 12 yaitu kriteria utama yang digunakan dalam proses klasifikasi yaitu nilai Matematika. Nilai matematika berfungsi sebagai *root* dari pohon keputusan yang dihasilkan. Kriteria pendukung nilai matematika dalam klasifikasi jurusan yaitu penghasilan orang tua.

3.3 Pengujian Algoritma

Pengujian dalam penelitian dilakuan dengan 2 cara yaitu pengujan algotrima K-Means dengan menghitung persentase dari setiap iterasi menggunakan *K-Fold Cross Validation* dan pengujian Algoritma C4.5 dengan nilai *Accuracy*, *Precision*, dan *Recall*. Sebelum menghitung nilai akurasi dari algoritma K-Means, terlbih dahulu menghitung jumlah prediksi benar dan prediksi salah yan g dihasilkan oleh proses algoritma K-means tersebut. Hasil perhitungan prediksi benar salah ditunjukkan Tabel 4.

Tabel 4. Prediksi benar salah dari hasil algoritma K-Means

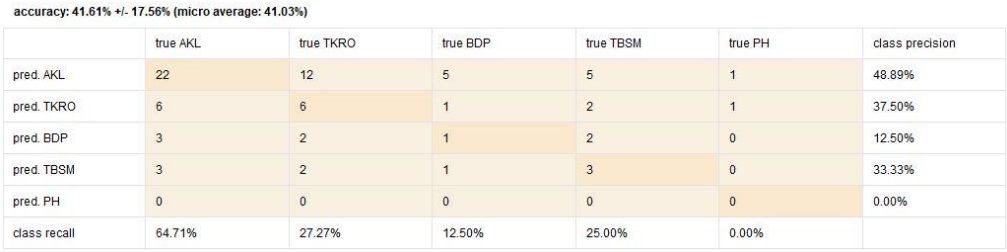
Minat	Hasil Algoritma K-Means	Prediksi	
		Benar	Salah
AKL	AKL	1	
TKRO	PH		1
AKL	AKL	1	
AKL	PH		1
AKL	BDP		1
TKRO	PH		1
TKRO	PH		1
TKRO	PH		1

TKRO	PH		1
AKL	AKL	1	
AKL	PH		1
AKL	AKL	1	
……	……	…….	…….
AKL	AKL	1	
BDP	BDP	1	
TKRO	TKRO	1	
Jumlah		**46**	**33**

Setelah mendapatkan nilai prediksi benar salah yang telah ditunjukkan pada Tabel 4, maka langkah selanjutnya yaitu menghitung nilai akurasi dari algoritma K-Means menggunakan *K-Fold Cross Validation* dan nilai *Precision*, *Recall*, dan *Accuracy* dari algoritma C4.5. Perhitungan nilai akurasi K-Means ditunjukkan pada Tabel 5, sedangkan hasil algoritma C4.5 ditunjukkan pada Gambar 13.

Tabel 5. Perhitungan Akurasi Algoritma K-Means

Jumlah Prediksi Benar	Jumlah Prediksi Salah	Perhitungan	Akurasi
46	33	$\frac{46}{79}\ x\ 100\%$	58,22%



accuracy: 41.61% +/- 17.56% (micro average: 41.03%)

	true AKL	true TKRO	true BDP	true TBSM	true PH	class precision
pred. AKL	22	12	5	5	1	48.89%
pred. TKRO	6	6	1	2	1	37.50%
pred. BDP	3	2	1	2	0	12.50%
pred. TBSM	3	2	1	3	0	33.33%
pred. PH	0	0	0	0	0	0.00%
class recall	64.71%	27.27%	12.50%	25.00%	0.00%	

Gambar 13. Nilai *Accuracy*, *Precision*, dan *Recall*

Precision $= \frac{0,4889+0,375+0,125+0,3333+0,00}{5} x100\% = 26,44\%$

Recall $= \frac{0,6471+0,2727+0,125+0,250+0,00}{5} x100\% = 25,9\%$

Accuracy $= \frac{32}{79} x100\% = 41,61\%$

Dari hasil pengujian akurasi algoritma K-Means yang ditunjukkan pada Tabel 5 didapatkan nilai akurasi sebesar 58,22%. Sedangkan pengujian *Accuracy*, *Precision*, dan *Recall* dari algoritma C4.5 yang ditunjukkan pada Gambar 5 didapatkan nilai *Accuracy* sebesar 41,61%, nilai *Precision* sebesar 26,44%, dan nilai *Recall* sebesar 25,9%. Kedua algoritma tersebut cukup baik untuk proses klasifikasi jurusan pada SMK.

4 PENUTUP

Setelah melalui beberapa tahap penelitian dalam proses penentuan jurusan siswa menggunakan Algoritma K-Means dan C4.5 diambil hasil akhir bahwa kedua Algoritma tersebut dapat digunakan untuk mengklasifikasikan jurusan bagi siswa SMK. Hasil pengklasifikasian menunjukkan bahwa variabel yang menjadi faktor utama dalam penentuan jurusan yaitu minat siswa karena minat siswa sebagai node atau akar lalu disusul dengan variabel nilai MTK. Penjurusan yang dilakukan dibagi menjadi 5 kategori jurusan yaitu Akutansi, Pemasaran, Perhotelan, TKR, dan TSM. Hasil pengujian kinerja klasifikasi penentuan jurusan menggunakan algoritma K-Means menghasilkan nilai akurasi sebesar 58,22%, sedangkan pengujian menggunakan algoritma C4.5 menghasilkan nilai *Precision* sebesar 26,44%, *Recall* sebesar 25,9%, dan *Accuracy* sebesar 41,61%,

Dilihat dari hasil pengujian tersebut disimpulkan bahwa algoritma K-Means memiliki tingkat akurasi yang lebih tinggi sehingga baik untuk melakukan klasifikasi jurusan bagi siswa SMK. Adapun beberapa saran untuk penelitian-penelitian selanjutnya yaitu dapat menerapkan algoritma lain sehingga didapatkan hasil yang dapat dibandingkan dengan penelitian sebelumnya. Penelitian selanjutnya dapat juga menggunakan kriteria lain dalam proses klasifikasi jurusan hingga mendapatkan akurasi yang lebih tinggi dan hasil pengklasifikasian yang lebih baik.

DAFTAR PUSTAKA

Bustami. (2010). Penerapan Algoritma Naive Bayes untuk Mengklasifikasi Data Nasabah. *TECHSI: Jurnal Penelitian Teknik Informatika*, *4*, 127–146.

Cerah, T. P. N., Nurhayati, O. D., & Isnanto, R. R. (2019). Perbandingan Metode Segmentasi K-Means Clustering dan Segmentasi Region Growing untuk Pengukuran Luas Wilayah Hutan Mangrove. *Jurnal Teknologi Dan Sistem Komputer*, *7*(1), 31–37.

Dharmayanti, D., Bachtiar, A. M., & Prasetyo, A. C. (2017). Penerapan Metode Clustering Untuk Membentuk Kelompok Belajar Menggunakan Di Smpn 19 Bandung. *Jurnal Ilmiah Komputer Dan Informatika*, *6*(2), 49–56.

Faisal, S. (2019). Klasifikasi Data Minning Menggunakan Algoritma C4.5 Terhadap Kepuasan Pelanggan Sewa Kamera Cikarang. *Jurnal Ilmu Komputer & Teknologi Informasi*, *4*(March), 38–45.

Firza, F., & Sarjono, S. (2020). Penerapan Algoritma K-Means Dalam Metode Clustering Untuk Peminatan Jurusan Bagi Siswa Swasta Pelita Raya Kota

Jambi. *Jurnal Manajemen Sistem Informasi*, *5*(3), 371–382.

Jadhav, S. D., & Channe, H. P. (2016). Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques. *International Journal of Science and Research*, *5*(1), 1842–1845.

Kantardzic, M. (2011). *Data mining: concepts, models, methods, and algorithms*. John Wiley & Sons.

Kurniasari, R., & Fatmawati, A. (2019). Penerapan Algoritma C4.5 Untuk Penjurusan Siswa Sekolah Menengah Atas. *Komputa : Jurnal Ilmiah Komputer Dan Informatika*, *8*(1), 19–27. https://doi.org/10.34010/komputa.v8i1.3045

Mardiani, M. (2015). Perbandingan Algoritma K-Means dan EM untuk Clusterisasi Nilai Mahasiswa Berdasarkan Asal Sekolah. *Creative Information Technology Journal*, *1*(4), 316.

Mohankumar, M., Amuthakkani, S., & Jeyamala, G. (2016). Comparative Analysis Of Decision Tree Algorithms For The Prediction Of Eligibility Of A Man For Availing Bank Loan. *International Journal of Advanced Research in Biology Engineering Science and Technology*, *2*(15), 360–366.

Nasari, F., & Darma, S. (2015). Seminar Nasional Teknologi Informasi dan Multimedia 2015 Penerapan K-Means Clustering Pada Data Penerimaan Mahasiswa Baru (Studi Kasus : Universitas Potensi Utama). *Seminar Nasional Teknologi Informasi Dan Multimedia*, 6–8.

Nur, F., Zarlis, M., & Nasution, B. B. (2017). Penerapan Algoritma K-Means Pada Siswa Baru Sekolahmenengah Kejuruan Untuk Clustering Jurusan. *Jurnal Nasional Informatika Dan Teknologi Jaringan*, *1*(2), 100–105.

Ong, J. O. (2013). Implementasi Algotritma K-means clustering untuk menentukan strategi marketing president university. *Jurnal Ilmiah Teknik Industri*, *12*(juni), 10–20.

Oscario, O., Jasmir, J., & Novianto, Y. (2019). Penerapan Algoritma C4.5 Untuk Memprediksi Kecocokan Gaya Belajar Bagi Siswa Siswi Sekolah Dasar (Studi Kasus : SD Sariputra Jambi). *Jurnal Processor*, *14*(2), 141.

Prasatya, A., Siregar, R. R. A., & Arianto, R. (2020). Penerapan Metode K-Means Dan C4.5 Untuk Prediksi Penderita Diabetes. *Jurnal Pengkajian Dan Penerapan Teknik Informatika*, *13*(1), 86–100. https://doi.org/10.33322/petir.v13i1.925

Putri, R. P. S., & Waspada, I. (2018). Penerapan Algoritma C4.5 pada Aplikasi Prediksi Kelulusan Mahasiswa Prodi Informatika. *Khazanah Informatika: Jurnal Ilmu Komputer Dan Informatika*, *4*(1), 1.

Tan, Steinbach, & Kumar. (2016). *Introduction to Data Mining*. United States of America: Pearson Education Limited.

Tandy, J., & Assegaff, S. (2019). Analisis Dan Perancangan Clustering Siswa Baru

Menggunakan Metode K-Means. *Manajemen Sistem Informasi*, *4*(4), 389–399.

Umar, R., Fadlil, A., & Az-Zahra, R. R. (2017). Pengelompokan Peminatan Jurusan di SMK Menggunakan Metode Self Organizing Map (SOM). *Seminar Nasional Teknologi Dan Komunikasi*, 203–210.

Wahyuni, S. (2018). Implementation of Data Mining to Analyze Drug Cases Using C4.5 Decision Tree. *Journal of Physics: Conference Series*, *970*(1), 0–6.