APPLICATION OF THE K-MEANS CLUSTERING METHOD FOR SELECTION OF INDUSTRIAL INTERNSHIP PROGRAM STUDENTS (CASE STUDY OF TUNAS YOUTH VOCATIONAL SCHOOL, TANGERANG)

Anton Susilo[1] Rizka Desiana[2]

anton.susilo@raharja.info[1], rizka.desiana@raharja.info[2]

Abstrak

Tunas Pemuda Tangerang Vocational School is a private vocational high school located in Tangerang Regency, Banten Province, which has quite a large number of students. Every year, Tunas Pemuda Tangerang Vocational School must place students for internship programs in industry. The large number of students who must be placed according to the needs of the industry which has collaborated with SMK Tunas Pemuda Tangerang, results in difficulties in determining which students are selected and grouped according to the students' abilities and the needs of the industry. Therefore, the aim of this research is to formulate and analyze problems, so that these problems can be overcome using the K-Means Clustering Algorithm method based on grouping data on report card grades in selected subjects and grades that have been determined by SMK Tunas Pemuda Tangerang as variables that will be supporting value in the selection of students for industrial internship programs. Students are selected based on their learning abilities in 8 selected subjects, namely PAI, PKN, Indonesian, Mathematics, English, Skills Competency 1, Skills Competency 2, and Skills Competency 3. Students who have different knowledge and skills will be grouped using K -Means Clustering into 3 categories using excel calculations and rapidminer tools. The conclusion of the results of this research is that from this data it was found that students with the knowledge and skill scores determined by the Vocational School would be grouped together according to the criteria so that teachers could group these students. From this method, a faster and more accurate grouping pattern is obtained according to what SMK and DUDI need. The attributes used are nis, student name and knowledge value, attendance and extracurriculars on report cards for grades 10, 11 and 12 semester 2. The K-Means Clustering Algorithm method is used to process these attributes to produce 3 clusters of students for the industrial internship program.

Keywords: K-Means Clustering, student selection, industrial internship program, rapidminer.

I. INTRODUCTION

SMK adalah lembaga pendidikan dibawah naungan Kementerian Pendidikan yang merupakan sekolah dengan basicdominan keterampilan dan keahlian, sehingga harus terus berinovasi untuk meningkatkan kualitas dibidang pendidikan kejuruan. Kegiatan belajar mengajar merupakan interaksi timbal balik dari guru dan juga siswa [1]. Untuk meningkatkan kualitas pendidikan, SMK Tunas Pemuda Tangerang mengakomodir potensi siswa dalam bidang akademik melalui program magang industri. Langkah-langkah menentukan siswa masuk ke program magang industri yang dilakukan olehSMK Tunas Pemuda saat ini di dapat dari hasil legger nilai terdiri 2 aspek penilaian yaitu nilai pengetahuan dan keterampilan serta nilai yang sudah melewati cakupan tingkat akurasi nilai tertinggi dari standar batasan keriteria ketutansan minimum (KKM). secara garis besar memiliki persamaan nilai variasi kriteria jumlah yang cenderung setara antar individu siswa. Selain itu, penempatan jumlah siswa program magang industri setiap tahun ajaran dapat berubah-ubah seiring dengan pengembangan SMK dan jumlah penerimaan siswa baru. Pemecahan permasalahan pengelompokan siswa dengan data yang semakin banyak menjadi kurang efesien dan dibutuhkan pengelompokan siswa yang juga memiliki potensi akademik untuk program magang industri yang ditetapkan SMK, mengakibatkan kesulitan yang dihadapi madrasah dalam menentukan siswa untuk program magang industri sesuai dengan kemampuan yang dimiliki.

Salah satu cara untuk mengatasi masalah ini adalah dengan clustering data yang bisa digunakan untuk pengolahan

data menjadi sumber informasi strategis [2] kemudian dikelompokan kedalam beberapa cluster berdasarkan kemiripan dari data-data tersebut [3]. Algoritma k-means clustering merupakan metode tertua [4] yang banyak diadopsi dalam penelitian [5] dan efektif dalam analisis cluster [6]. Beberapa penelitian sebelumnya, telah berhasil melakukan pengelompokan dengan menggunakan algoritma K-Means, diantaranya adalah: Guiyun Feng, et.al [7] Hasil eksperimen menunjukkan bahwa statistik tidak hanya memecahkan kesulitan untuk menentukan jumlah pengelompokan dalam algoritma K-means dari sudut pandang objektif dan kuantitatif, tetapi juga meningkatkan keandalan hasil prediksi. Qin-ZhuoLiao, et.al [8] pengelompokan K-means untuk memilih hanya beberapa perwakilan, di mana dua fase simulasi aliran diimplementasikan. Model empiris kemudian diadopsi untuk menggambarkan hubungan antara solusi fase tunggal dan solusi dua fase menggunakan perwakilan ini. Hasilnya dua fase di semua realisasi dapat diprediksi menggunakan model empiris dengan mudah. M Mughnyanti, et.al [9]. Studi penelitian ini menggunakan algoritma K-Means dengan evaluasi Indeks Davies-Bouldin untuk menentukan jumlah cluster Centroid dilakukan dengan memodifikasi K-Means metode untuk melakukan beberapa penentuan centroid untuk mendapatkan Iterasi. Hasilnya adalah menghasilkan anggota cluster yang memiliki tingkat kemiripan yang baik dengan data lainnya. Mochammad Faid, et.al [10], Dalam penelitian ini mencoba untuk melihat kinerja algoritma data mining dengan menggunakan tool data mining. Adapun tool data mining yang akan digunakan adalah Microsoft Excel dan Rapidminer. penelitian ini mencoba mengkomparasikan tool data mining yang sering digunakan oleh para peneliti yaitu Rapidminer dan Excel. Nirsal, et.al [11] Penelitian ini bertujuan mendesain dan mengimplementasikan sistem pembelajaran berbasis elearning pada Sekolah Menengah Kejuruan dengan software pembangun aplikasi menggunakan balsamic mockup. Dalam penelitian ini, penulis menggunakan metode algoritma K-Means Clustering, dengan perbandingan perhitungan konvensional menggunakan microsoft excel dan menggunakan tool rapidminer, serta mengimplementasikannya dengan aplikasi K-Means Clustering Seleksi program magang industri. Tool RapidMiner digunakan untuk melakukan analisis prediktif dan penambangan data [12], untuk menerapkan beberapa algoritma pembelajaran mesin, termasuk pohon keputusan [13]. Penulis akan mengelompokkan siswa yang lolos seleksi akan menempati kelas unggulan sedangkan siswa yang tidak lolos seleksi akan masuk ke kelas reguler. Data yang digunakan peneliti ini berupa kumpulan nilai rapor kelas 10, 11 dan 12 semester 2 yang diambil dari data Wakil Kepala SMK bidang kurikulum. Kumpulan data ini terdiri dari nilai 8 mata pelajaran yang telah ditentukan oleh pihak madrasah, absensi kehadiran dan ekstrakurikuler serta memiliki 3 cluster yang digunakan dalam menentukan kelas siswa yang lolos seleksi program magang industri. Berdasarkan latar belakang yang telah dijabarkan di atas, penelitian ini menjawab bagaimana caranya mengelompokkan siswa untuk program magang industri dengan menggunakan metode k-means clustering di SMK Tunas Pemuda dan bagaimana menentukan hasil clustering untuk data siswa program magang industri dengan tujuan penelitian adalah untuk mengetahui hasil pengelompokkan siswa untuk program magang industri, berdasarkan data siswa dengan menggunakan metode k-means clustering di SMK Tunas Pemuda Tangerang dan bisa memberikan langkah strategis bagi SMK Tunas Pemuda Tangerang dalam menyeleksi siswa dengan metode K-Means Clustering.

2. MATERIALS AND METHODS

Bagian ini menjelaskan detail kumpulan data yang digunakan dan metode klasifikasi. Data dalam penelitian ini adalah data siswa kelas 10, 11 dan 12 sebanyak 24 rombel dengan total 741 siswa terdiri dari atribut nis, nama siswa, nilai mata pelajaran, kehadiran, ekstrakulikuler. Contoh dari beberapa data yang dikumpulkan dari siswa ditunjukkan pada Tabel 1.

Table 1. The examples of student data

No	Nisn/Nis	Name	knowledge subjects								attendance						extracurricular		
			PAI	PPKN	B.IND	MTK	B.ING	KK1	KK2	KK3	JAN	FEB	MAR	APR	MEI	JUN	Wajib	Pilihan	Keagamaan
1	0078862723 / 22231106	AIRA RAMANIYA	88	86	89	90	91	93	89	90	2	0	1	0	1	0	1	1	1
2	3078771215 / 22231107	AMANDA AYU LESTARI	85	83	84	82	80	81	83	83	1	0	0	0	1	1	1	0	1
3	0079600442 / 22231108	ANDRE KURNIAWAN	86	85	87	90	92	92	93	90	1	1	1	1	1	1	1	1	1
4	0077045950 / 22231109	ARNELITA RETVI DAMAYANTI	86	85	87	88	92	91	90	92	1	1	0	0	0	1	1	1	1
5	0077256506 / 22231110	ASHER WIRYATEJA	89	87	88	87	91	89	90	91	2	2	0	0	1	0	1	1	1
6	0061170250 / 22231111	Audy Christy Wijaya	83	82	82	83	81	81	83	80	0	0	0	0	0	2	1	1	1
7	3073090064 / 22231112	AULIA	77	79	80	78	81	81	78	76	1	0	1	1	0	0	1	1	0
8	0069530094 / 22231113	Citra Rizky Ramadhan	83	82	84	82	80	82	82	80	0	2	0	1	1	0	1	0	1
9	0063918566 / 22231114	Decha Fatimahtuz Zahra	87	86	90	91	90	91	89	90	1	0	0	0	1	1	1	1	0
10	0084513252 / 22231115	DENY SETIAWAN	80	81	81	82	84	80	81	82	2	0	0	0	0	0	1	1	1
11	0076327361 / 22231116	DIMAS OKA AGUSTIAN	87	86	88	89	92	90	92	90	0	1	1	0	1	1	1	1	1
12	0071820027 / 22231117	DINI ANGGREYANI	84	83	83	82	81	80	82	83	1	1	0	1	0	0	1	1	1
13	0079078946 / 22231278	ECIN CLAUDIA	86	88	87	90	91	93	91	89	2	0	0	0	1	0	1	0	1
14	0066963347 / 22231118	ELVINA DWI ASTUTI	79	80	77	80	78	81	79	78	0	1	0	2	0	1	1	1	1
15	0077852524 / 22231119	FAJAR ALQORI	80	81	80	84	88	88	82	85	0	0	0	0	1	0	1	1	1
16	0076707520 / 22231120	Firda Zahratul Syifa	82	82	84	80	79	81	79	82	0	1	0	0	1	1	1	1	0
17	0061537571 / 22231121	Helda Agustina	79	80	77	84	84	78	77	82	0	0	0	1	0	1	1	1	1
18	0096095261 / 22231122	Imel Sagita	77	80	79	78	82	82	81	80	1	0	1	0	0	0	1	1	1
19	0061506583 / 22231123	MUHAMMAD FIKRI FADILAH	84	83	80	81	85	85	84	80	0	1	0	0	0	1	1	0	1
20	0066739756 / 22231124	MUHAMMAD SOPIANA	85	86	87	88	89	91	88	87	0	0	0	1	0	0	1	0	1

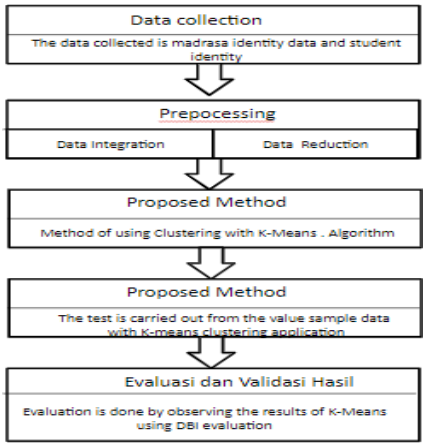

Figure 1. Description of the process in the K-means Clustering method

2.1 Pre-processing

Preprocessing data adalah tahapan dari Data Mining yaitu suatu proses atau tahapan yang dilakukan untuk mengolah data mentah menjadi data yang berkualitas atau inputan yang baik untuk dilanjutkan ke proses selanjutnya. Tahap preprocessing data ini adalah tahap yang sangat krusial atau sangat penting dan harus dilakukan dengan teliti karena proses Data Mining membutuhkan data yang konsisten dalam penulisan, benar dalam formatnya, tidak ada data yang kosong, duplikasi data, dan lain lain. Data yang tidak berkualitas maka hasil dari proses Data Mining ini akan menghasilkan hasil yang tidak berkualitas juga. Dari dataset nilai pelajaran, 15 nilai mata pelajaran dipilih menjadi 8 nilai mata pelajaran sesuai dengan kriteria ujian nasional SMK dan kompetensi sains SMK. Jumlah keseluruhan nilai masing-masing data set mata pelajaran, absensi kehadiran dan ekstrakurikuler diinterval menjadi satu sampai empat.

Table 2. Pre-processing results

No	Nisn/Nis	Name	subject vslue	attendance	Extra curiculer
1	0078862723 / 22231106	AIRA RAMANIYA	4	3	3
2	3078771215 / 22231107	AMANDA AYU LESTARI	2	4	2
3	0079600442 / 22231108	ANDRE KURNIAWAN	4	3	3
4	0077045950 / 22231109	ARNELITA RETVI DAMAYANTI	4	4	3
5	0077256506 / 22231110	ASHER WIRYATEJA	4	3	3
6	0061170250 / 22231111	Audy Christy Wijaya	2	4	3
7	3073090064 / 22231112	AULIA	1	4	2
8	0069530094 / 22231113	Citra Rizky Ramadhan	2	3	2
9	0063918566 / 22231114	Decha Fatimahtuz Zahra	4	4	2
10	0084513252 / 22231115	DENY SETIAWAN	2	4	3
….					
741	0030062136 / 20211033	Tomi Wijaya	1	4	2

2.2 Classification

Proses klasifikasi dilakukan dengan menggunakan K-Means Clustering. K-Means adalah teknik pengelompokan data atau informasi menjadi kelompok untuk mendapatkan jumlah cluster yang tepat, yang dilakukan oleh menentukan nilai k sebelumnya. Semakin dekat jaraknya nilai menyimpulkan tingkat kesamaan yang lebih tinggi. Semakin tinggi nilai jarak, semakin tinggi ketidaksamaan tersebut.

2.2.1 K-Means Clustering

Dalam penerapan k-mean dengan sumber data dari rapor adalah sebagai berikut;

1. Memasukkan data awal
2. Menentukan total cluster yang diinginkan
3. Menentukan k centroid (titik pusat cluster) awal
4. Menghitung jarak terdekat ke centroid

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

5. Menghitung pusat cluster dengan anggota klaster yang baru
6. Ulangi langkah 3 – 5 hinga sudah tidak ada lagi data yang berpindah ke cluster lain.

2.2.2 Performance evaluation

Setelah proses perhitungan k-means secara konvensional menggunakan microsoft excel telah selesai dilakukan, selanjutnya dievaluasi proses perhitungan k-means dengan pengujian Davies Bouldin Index (DBI) secara konvensional menggunakan microsoft excel. Tahapan dari perhitungan Davies Bouldin Index adalah sebagai berikut:

1. Menghitung Sum of Square Within-cluster (SSW)

$$SSW_i = \frac{1}{m_i} \sum_{j=i}^{m_i} d(x_j, c_i)$$

2. Sum of Square Between-cluster (SSB)

$$SSB_{i,j} = d(c_i, c_j)$$

3. Setelah nilai kohesi dan separasi diperoleh, kemudian dilakukan pengukuran rasio (Rij

$$R_{i,j} = \frac{SSW_I + SSW_J}{SSB_{i,j}}$$

4. Nilai rasio yang diperoleh tersebut digunakan untuk mencari nilai davies bouldin index (DBI) dari persamaan berikut;

$$DBI = \frac{1}{k} \sum_{i=1}^{k} max_{i \neq j} (R_{i,j})$$

Dari hasil perhitungan DBI diatas, peneliti mengambil hasil DBI yang bernilai 0,729037 karena nilainya yang paling mendekati angka 0. Semakin kecil nilai DBI maka semakin baik. Semakin kecil nilai DBI yang diperoleh (non-negatif >= 0), maka semakin baik cluster yang diperoleh dari pengelompokan K-means yang digunakan.

2.2.2 Rapidminer

RapidMiner digunakan untuk melakukan analisis prediktif dan penambangan data, hal ini memungkinkan seseorang untuk menerapkan beberapa algoritma pembelajaran mesin, termasuk pohon keputusan

3. RESULTS AND DISCUSSION

Dalam penerapan k-mean dengan sumber data dari rapor adalah sebagai berikut:

1. Memasukkan data awal

a. Pengumpulan Data Data-data yang diperoleh pada tahapan pengumpulan data menghasilkan 881 data siswa darirapor dengan atribut Nilai PAI, nlai PKN, nilai Bahasa Indonesia, nilai Matematika, nilai Bahasa Inggris, Nilai Kompetensi Keahlian 1, Nilai Kompetensi Keahlian 2, Nilai Kompetensi Keahlian 3, Nilai Absensi dan Ekstrakurikuler yang nantinya digunakan sebagai data perhitungan k-means. Berikut potongan hasil

preprocessing yang disajikan pada tabel berikut;

Table 3. Preprocessing

No	Nisn/Nis	Name	subject vslue	attendance	Extra curiculer
1	0078862723 / 22231106	AIRA RAMANIYA	4	3	3
2	3078771215 / 22231107	AMANDA AYU LESTARI	2	4	2
3	0079600442 / 22231108	ANDRE KURNIAWAN	4	3	3
4	0077045950 / 22231109	ARNELITA RETVI DAMAYANTI	4	4	3
5	0077256506 / 22231110	ASHER WIRYATEJA	4	3	3
6	0061170250 / 22231111	Audy Christy Wijaya	2	4	3
7	3073090064 / 22231112	AULIA	1	4	2
8	0069530094 / 22231113	Citra Rizky Ramadhan	2	3	2
9	0063918566 / 22231114	Decha Fatimahtuz Zahra	4	4	2
10	0084513252 / 22231115	DENY SETIAWAN	2	4	3
....					
741	0030062136 / 20211033	Tomi Wijaya	1	4	2

b. Menentukan total cluster yang diinginkan Dalam menentukan total cluster dari data-data yang ada akan dibuat menjadi 3 cluster.

c. Menentukan k centroid (titik pusat cluster)

Centroid awal akan dipilih pada data acak yang telah ditentukan. Data diambil pada data. Untuk pusat cluster ke-1 di ambil data ke-104, cluster ke-2 di ambil data ke-234, dan cluster ke-3 di ambil data ke-723. Data akan disajikan dalam tabel 4

Table 4. The first Centroid

	X	Y	Z
Data Ke-104	2	4	3
Data Ke-234	4	3	2
Data Ke-723	1	3	2

d. Menghitung jarak terdekat ke centroid

Berikut hasil perhitungan jarak setiap data ke masing-masing centroid tersaji dalam tabel 5

Table 5 Centroid Distance Results

No	NISN / NIS	Name	Subject Value	Attendance	Extra	C1	C2	C3	Nearest Distance	Cluster
1	0078862723 / 22231106	AIRA RAMANIYA	4	3	3	2,2361	1,0000	3,1623	1,0000	2
2	3078771215 / 22231107	AMANDA AYU LESTARI	2	4	2	1,0000	2,2361	1,4142	1,0000	1
3	0079600442 / 22231108	ANDRE KURNIAWAN	4	3	3	2,2361	1,0000	3,1623	1,0000	2
4	0077045950 / 22231109	ARNELITA RETVI DAMAYANTI	4	4	3	2,0000	1,4142	3,3166	1,4142	2
5	0077256506 / 22231110	ASHER WIRYATEJA	4	3	3	2,2361	1,0000	3,1623	1,0000	2
6	0061170250 / 22231111	Audy Christy Wijaya	2	4	3	0,0000	2,4495	1,7321	0,0000	1
7	3073090064 / 22231112	AULIA	1	4	2	1,4142	3,1623	1,0000	1,0000	3
8	0069530094 / 22231113	Citra Rizky Ramadhan	2	3	2	1,4142	2,0000	1,0000	1,0000	3
9	0063918566 / 22231114	Decha Fatimahtuz Zahra	4	4	2	2,2361	1,0000	3,1623	1,0000	2
10	0084513252 / 22231115	DENY SETIAWAN	2	4	3	0,0000	2,4495	1,7321	0,0000	1

Data tabel 5 hasil perhitungan iterasi ke 1, selanjutnya dikelompokkan dengan 3 cluster seperti tabel 6:

Table 6 Grouping of Data in Iteration 1

Literation 1			
Cluster Center			
C1	2	4	2
C2	4	3	3
C3	1	3	3
Name	**C1**	**C2**	**C3**
AIRA RAMANIYA	2,2361	1,0000	3,1623
AMANDA AYU LESTARI	1,0000	2,2361	1,4142
ANDRE KURNIAWAN	2,2361	1,0000	3,1623
ARNELITA RETVI DAMAYANTI	2,0000	1,4142	3,3166
ASHER WIRYATEJA	2,2361	1,0000	3,1623
Audy Christy Wijaya	0,0000	2,4495	1,7321
AULIA	1,4142	3,1623	1,0000
Citra Rizky Ramadhan	1,4142	2,0000	1,0000
Decha Fatimahtuz Zahra	2,2361	1,0000	3,1623
DENY SETIAWAN	0,0000	2,4495	1,7321

e. Menghitung pusat klaster dengan anggota klaster yang baru
Setelah didapatkan hasil iterasi 1, maka dilakukan penghitungan pusat cluster baru untuk iterasi selanjutnya.

Table 7. New Cluster Deteemination from iteration 1

Penentuan Cluster Baru Dari Iterasi Ke-1			
C1	2	4	3
C2	4	3	3
C3	1	3	2

f. Proses klaster sudah selesai bila pusat klaster tidak berubah, namun jika pusat klaster masih berubah maka diulangi langkah menghitung jarak hingga pusat klaster tidak berubah lagi.

Table 8 The result C1,C2,C3 Iterasi 1, Iterasi 2, Iterasi 3, Iterasi 4, dan Iterasi 5

Iteration 1			
Cluster Center			
C1	2	4	2
C2	4	3	3
C3	1	3	3
Name	**C1**	**C2**	**C3**
AIRA RAMANIYA	2,2361	1,0000	3,1623
AMANDA AYU LESTARI	1,0000	2,2361	1,4142
ANDRE KURNIAWAN	2,2361	1,0000	3,1623
ARNELITA RETVI DAMAYANTI	2,0000	1,4142	3,3166
ASHER WIRYATEJA	2,2361	1,0000	3,1623
Audy Christy Wijaya	0,0000	2,4495	1,7321
AULIA	1,4142	3,1623	1,0000
Citra Rizky Ramadhan	1,4142	2,0000	1,0000
Decha Fatimahtuz Zahra	2,2361	1,0000	3,1623
DENY SETIAWAN	0,0000	2,4495	1,7321

Iteration 2			
Cluster Center			
C1	2	4	3
C2	4	3	3
C3	1	3	2
Name	**C1**	**C2**	**C3**
AIRA RAMANIYA	2,2773	0,4113	2,6296
AMANDA AYU LESTARI	1,0223	2,2252	1,7152
ANDRE KURNIAWAN	2,2773	0,4113	2,6296
ARNELITA RETVI DAMAYANTI	2,2155	0,9359	2,8476
ASHER WIRYATEJA	2,2773	0,4113	2,6296
Audy Christy Wijaya	0,4105	2,1716	1,2718
AULIA	1,2943	3,1419	1,6420
Citra Rizky Ramadhan	2,4053	2,0603	1,3222
Decha Fatimahtuz Zahra	2,4053	1,0542	3,0714
DENY SETIAWAN	0,4105	2,1716	1,2718

Iteration 3			
Cluster Center			
C1	2	4	3
C2	4	3	3
C3	1	3	3
Name	**C1**	**C2**	**C3**
AIRA RAMANIYA	2,4735	0,4215	2,6371
AMANDA AYU LESTARI	0,6270	2,1453	1,5605
ANDRE KURNIAWAN	2,4735	0,4215	2,6371
ARNELITA RETVI DAMAYANTI	2,3467	0,9401	2,8892
ASHER WIRYATEJA	2,4735	0,4215	2,6371
Audy Christy Wijaya	0,5920	2,0845	1,3650
AULIA	0,9027	3,0550	1,4809
Citra Rizky Ramadhan	1,0022	1,9738	1,0208
Decha Fatimahtuz Zahra	2,3557	1,0682	2,9865
DENY SETIAWAN	0,5920	2,0845	1,3650

Iteration 4			
Cluster Center			
C1	2	4	2
C2	4	3	3
C3	1	3	3
Name	**C1**	**C2**	**C3**
AIRA RAMANIYA	2,5494	0,4231	2,6401
AMANDA AYU LESTARI	0,6312	2,1119	1,5621
ANDRE KURNIAWAN	2,5494	0,4231	2,6401
ARNELITA RETVI DAMAYANTI	2,4297	0,8923	2,8926
ASHER WIRYATEJA	2,5494	0,4231	2,6401
Audy Christy Wijaya	0,6700	2,0347	1,3683
AULIA	0,8192	3,0219	1,4808
Citra Rizky Ramadhan	0,9972	1,9604	1,0213
Decha Fatimahtuz Zahra	2,4193	1,0567	2,9892
DENY SETIAWAN	0,6700	2,0347	1,3683

Iteration 5			
Cluster Center			
C1	2	4	2
C2	4	3	3
C3	1	3	3
Name	**C1**	**C2**	**C3**
AIRA RAMANIYA	2,5494	0,4231	2,6401
AMANDA AYU LESTARI	0,6312	2,1119	1,5621
ANDRE KURNIAWAN	2,5494	0,4231	2,6401
ARNELITA RETVI DAMAYANTI	2,4297	0,8923	2,8926
ASHER WIRYATEJA	2,5494	0,4231	2,6401
Audy Christy Wijaya	0,6700	2,0347	1,3683
AULIA	0,8192	3,0219	1,4808
Citra Rizky Ramadhan	0,9972	1,9604	1,0213
Decha Fatimahtuz Zahra	2,4193	1,0567	2,9892
DENY SETIAWAN	0,6700	2,0347	1,3683

Dari tabel diketahui iterasi dilakukan sebanyak empat kali dengan jumlah akhir cluster C1, C2 dan C3 iterasi lima sebagai berikut :

Table 9 Result C1,C2,C3 Iterasi 5

	C1	207
	C2	257
	C3	277
Jumlah		**741**

2. Perhitungan Evaluasi Clustering

Evaluasi proses perhitungan k-means dengan pengujian Davies Bouldin Index (DBI) secara konvensional menggunakan microsoft excel. Tahapan dari perhitungan Davies Bouldin Index adalah sebagai berikut:

a. Sum of Square Within-cluster (SSW)

Untuk mengetahui kohesi dalam sebuah cluster ke-i adalah denganmenghitung nilai dari Sum of Square Within-cluster (SSW).

Table 10. Last value of K-Means centroid

Centroid	X	Y	Z
C1	2	4	2
C2	4	3	3
C3	1	3	3

Kemudian menghitung jarak antar cluster pada anggota cluster dengan persamaan seperti dibawah ini:

$$\textbf{Jarak d,c} = \sqrt{(x_{1i} - x_{1j})^2 + (x_{2i} - x_{2j})^2 + (x_{3i} - x_{3j})^2 +}$$

$Dc_1 = \sqrt{(3 - 4)^2 + (0 - 3)^2 + (3 - 3)^2} = 2{,}9789$

$Dc_2 = \sqrt{(1 - 4)^2 + (2 - 3)^2 + (2 - 2)^2} = 2{,}3728$

$Dc_3 = \sqrt{(3 - 4)^2 + (0 - 3)^2 + (3 - 3)^2} = 2{,}9789$

Kemudian langkah selanjutnya menghitung Sum of Square Within cluster (SSW) dengan persamaan dibawah ini:

$$\textbf{SSW} = \frac{Jumlah\ hasil\ jarak\ cluster}{Jumlah\ anggota\ cluster}$$

$SSW_1 = \frac{2{,}3728+2{,}4557+1{,}6187+2{,}4746+\cdots+1{,}6187}{207} = 2{,}0154$

$SSW_2 = \frac{2{,}9789+2{,}9789+1{,}6305+2{,}9789+\cdots+2{,}9789}{257} = 2{,}7510$

$SSW_3 = \frac{2{,}5309+2{,}5693+2{,}5309+2{,}5309+\cdots+2{,}0637}{277} = 2{,}4365$

Hasil dari menghitung jarak antar anggota cluster dan SSW dapat dilihat pada tabel dibawah ini.

Table 11. The sample calculates the distance of Cluster 1 members as a result of Clustering k-means.

No	Name	C1	C2	C3	Cluster	Ctrd X	Ctrd Y	Ctrd Z	Nearest Distance	SSW
1	AIRA RAMANIYA	3	0	3	2	4	3	3	2,9789	2,7510
2	AMANDA AYU LESTARI	1	2	2	1	2	4	2	2,3728	2,0154
3	ANDRE KURNIAWAN	3	0	3	2	4	3	3	2,9789	2,4365
4	ARNELITA RETVI D	2	1	3	2	4	3	3	2,6305	
5	ASHER WIRYATEJA	3	0	3	2	4	3	3	2,9789	
6	Audy Christy Wijaya	1	2	1	1	2	4	2	2,4557	
7	AULIA	1	3	1	1	2	4	2	1,6187	
8	Citra Rizky Ramadhan	1	2	1	1	2	4	2	2,4746	
9	Decha Fatimahtuz Zahra	2	1	3	2	4	3	3	2,5051	
10	DENY SETIAWAN	1	2	1	1	2	4	2	2,4557	

b. Sum of Square Between-cluster (SSB)

Perhitungan Sum of Square Between-cluster (SSB) bertujuan untuk mengetahui separasi antar cluster Persamaan yang digunakan untuk menghitung nilai Sum of Square Between cluster adalah sebagai berikut:

$SSB_{i,j} = d(C_i.C_2)$

$= \sqrt{(C_i - C_j)}$

$SSB_{1,2} = \sqrt{(2{,}00 - 4{,}00)^2 + (3{,}7647 - 3{,}1977)^2 + (2{,}4118 - 3{,}0659)^2}$

$= 2{,}1793133$

$SSB_{1,3} = \sqrt{(2{,}00 - 1{,}00)^2 + (3{,}7647 - 2{,}8351)^2 + (2{,}4118 - 2{,}8351)^2}$

$= 1{,}4294594$

$SSB_{2,3} = \sqrt{(4{,}00 - 1{,}00)^2 + (3{,}1977 - 2{,}8351)^2 + (3{,}0659 - 2{,}8351)^2}$

$= 2{,}7957$

Table 12. Result SSB Algoritma K-Means

Count SSB	CENTROID		
	1	2	3
1	0	2,1793133	1,4294594
2	2,1793133	0	3,035595
3	1,4294594	3,0306261	0

c. pengukuran rasio (Rij)

Nilai rasio yang diperoleh tersebut digunakan untuk mencari nilai daviesbouldin index (DBI) dari persamaan berikut :

$$R_{i,j} = \frac{SSW_I + SSW_J}{SSB_{i,j}}$$ 109

$R_{13} = \frac{2{,}0154+2{,}4365}{1{,}4294594} = 3{,}1143942$

$R_{23} = \frac{2,7510+2,4365}{1,4294594} = 3,6289944$

Nilai rasio yang diperoleh tersebut digunakan untuk mencari nilai daviesbouldin index (DBI) dari persamaan berikut;

$$DBI = \frac{1}{k}\sum_{i=1}^{k} max_{i \neq j}\left(R_{i,j}\right)$$

$DBI_{12} = \frac{2,1871109}{3} = 0,729037$

$DBI_{13} = \frac{3,1143942}{3} = 1,0381314$

$DBI_{23} = \frac{3,6289944}{3} = 1,2096648$

Dari hasil perhitungan DBI diatas, peneliti mengambil hasil DBI yang bernilai 0.729037 karena nilainya yang paling mendekati angka 0. Semakin kecil nilai DBI maka semakin baik. Semakin kecil nilai DBI yang diperoleh (non-negatif >= 0), maka semakin baik cluster yang diperoleh dari pengelompokan K-means yang digunakan.
Dalam memvalidasi evaluasi DBI dengan cara konvensional (microsoft excel) pada penelitian ini, peneliti menggunakan Davies Bouldin Index (DBI) Rapidminer studio 10.2. Langkah awal adalah memasukan data seleksi siswa ke rapidminer studio 10.2 seperti pada gambar 2 dan parameter evaluasi menggunakan DBI dilihat pada gambar 3 dibawah ini.

Row No.	id	cluster	N. mata pelajaran	N. absensi	N. ekstrakurikuler
1	1	cluster_1	4	3	3
2	2	cluster_2	2	4	2
3	3	cluster_1	4	3	3
4	4	cluster_1	4	4	3
5	5	cluster_1	4	3	3
6	6	cluster_0	2	4	3
7	7	cluster_2	1	4	2
8	8	cluster_2	2	3	2
9	9	cluster_1	4	4	2
10	10	cluster_0	2	4	3

Figure 2. Initial data for student selection using K-Means Clustering

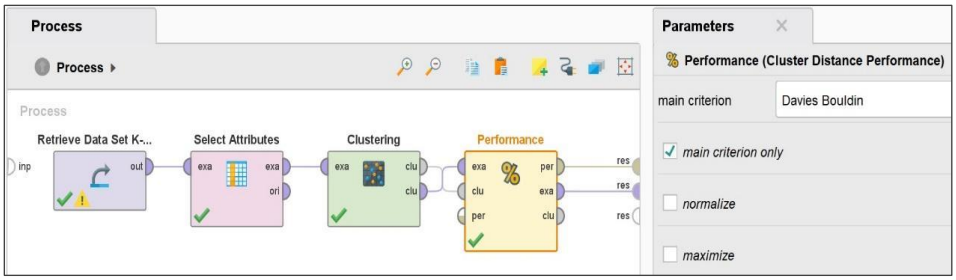

Figure 3 Parameters Method Davies Bouldin Index Clustering Algorithm.

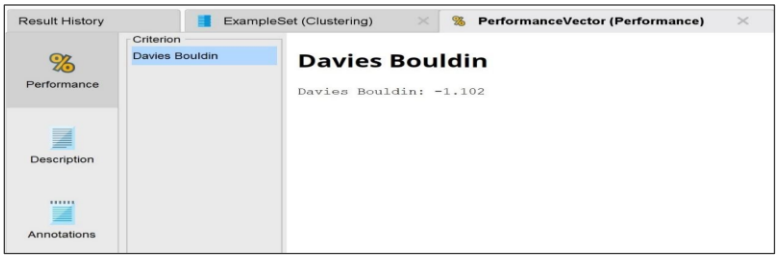

Figure 4. Davies Bouldin Index (DBI) K-Means Clustering Algorithm

Hasil dari nilai Davies Bouldin dan deskripsi Performance yang menghasilkan Centroid Distance pada setiap cluster yang ada didalam Performance Vector K-Means dapat dilihat pada gambar dibawah ini. Semakin rendah nilai DBI, maka cluster tersebut semakin baik.

Table 13. Standar Receiver Operating Characteristic (ROC)

Nilai Rasio	Kategori
0.80 - 1.00	Sangat Baik
0.60 - 0.80	Baik
0.40 - 0.60	Cukup Baik
0.20 - 0.40	Kurang Baik
0.00 - 0.20	Tidak Baik

3.1 Implementasi K-Means Clustering

Penerapan Metode *K-Means Clustering* dalam Seleksi Siswa Program Magang Industri (Studi Kasus SMK Tunas Pemuda Tangerang) ini diimplementasikan dengan menggunakan Balsamic Mockup dan dirancang agar mudah digunakan dalam menghasilkan seleksi siswa kelas unggulan. Berikut penjelasan interface dari prototype aplikasi perhitungan *K-Means Clustering*;

a. Tampilan Halaman Login

Halaman Login merupakan halaman pertama yang tampil pada saat link diakses. Di Halaman Login admin akan memasukan Username dan Password untuk masuk ke aplikasi, seperti yang ditunjukkan pada gambar 5

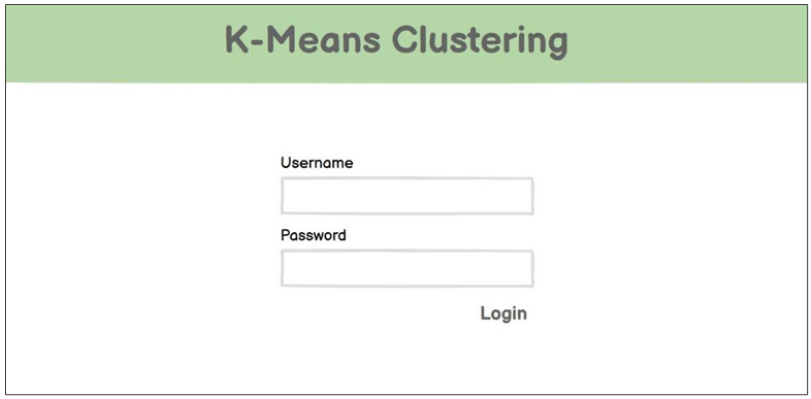

Figure 5 Login page display

b. Tampilan Halaman Utama

Dalam tampilan utama aplikasi ini terdapat 3 menu pilihan yang bisa digunakan dalam menghitung k-means clustering. Menu dalam tampilan utamanya adalah menu input data, iterasi data dan menu hasil clustering yang sudah diuji coba di awal menggunakan hitungan konvensional microsoft excel dan rapidminer 10.2

Figure 6. Main page display

c. Tampilan Halaman Input Data

Pada menu input data disediakan tool untuk tambah data secara mandiri, export xls dan bersihkan data. Untuk tool tambah data, data *entry* atau dimasukkan satu persatu sesuai dataset NIS, Nama Siswa dan penilaian siswa (mata pelajaran, absensi, dan ekstrakurikuler) tersaji pada gambar 7

Figure 7. Data Input Page Display

d. Tampilan Halaman Input Iterasi Data

Prototype aplikasi K-Mean Clustering ini terdapat menu untuk iterasi data dengan memilih data awal yang akan dijadikan pusat cluster untuk C1, C2 dan C3. Memilih nama siswa secara acak dan memasukan nilai mata pelajaran, prestasi dan ekstrakurikuler. Setelah C1, C2, dan C3 terisi seperti gambar 4.5, data tersebut bisa langsung diproses dan hasilnya dapat dilihat pada gambar 4.7. Data pada gambar 8, sudah diketahui C1, C2, dan C3 beserta pengelompokan datanya.

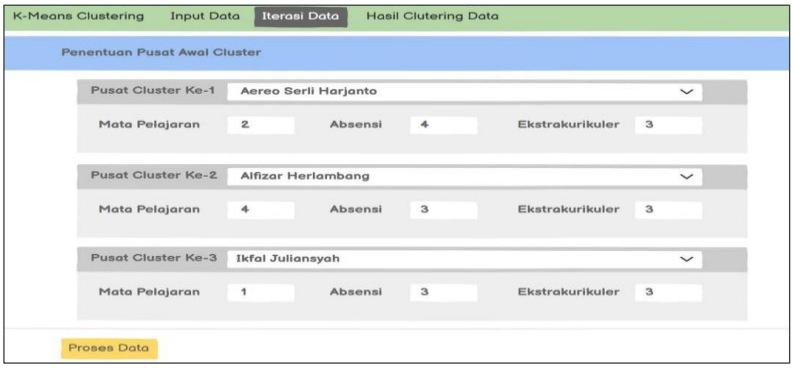

Figure 8 Display of Data Iteration Input Page

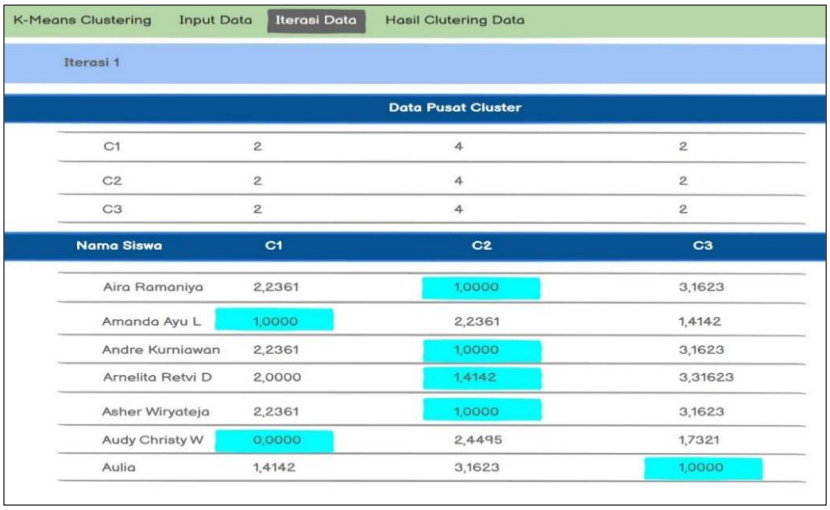

Figure 9. Display of Iteration 1 Results

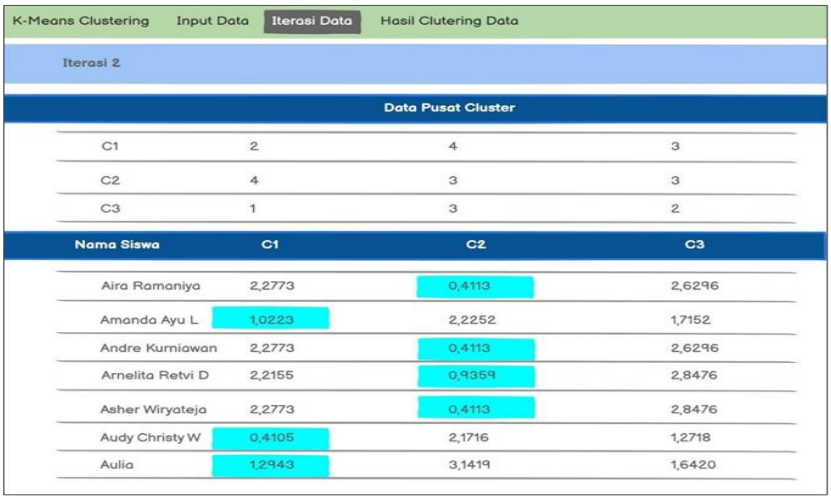

Figure 10. Display of Iteration 2 Results

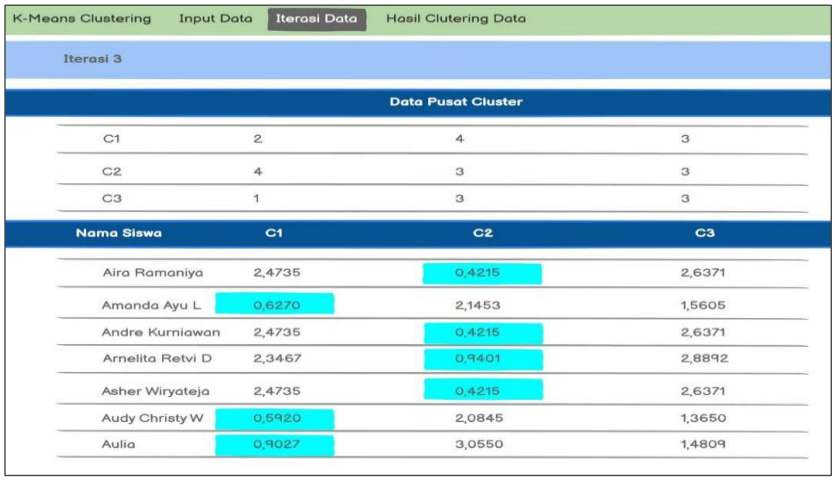

Figure 11. Display of Iteration 3 Results

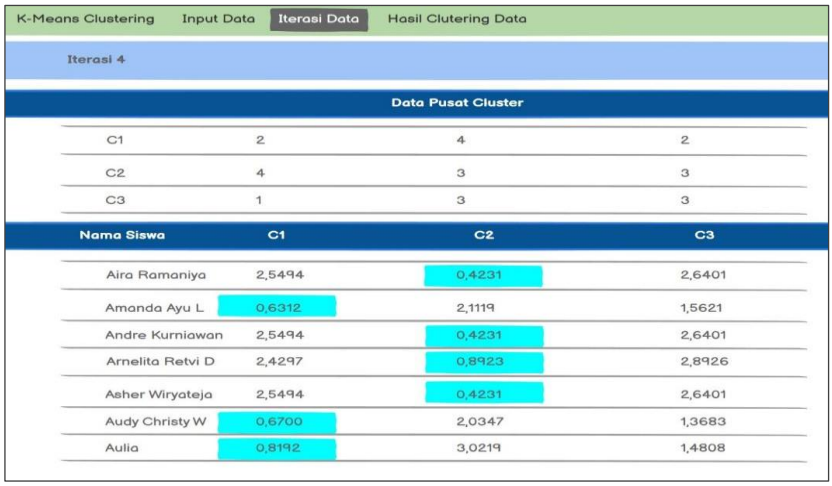

Figure 12. Display of Iteration 4 Results

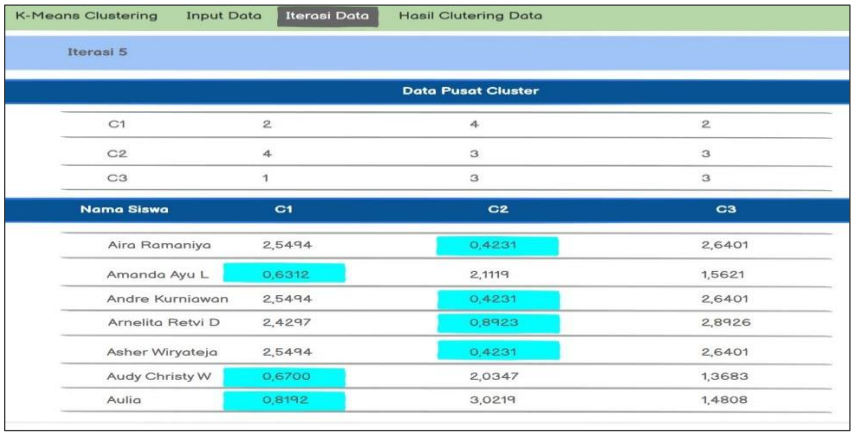

Figure 13. Display of Iteration 5 Results

e. Tampilan Halaman Hasil Clustering Data

Pada tampilan menu hasil clustering data, hasil yang disajikan adalah hasil iterasi, dan disajikan dalam bentuk persentase masing-masing cluster.

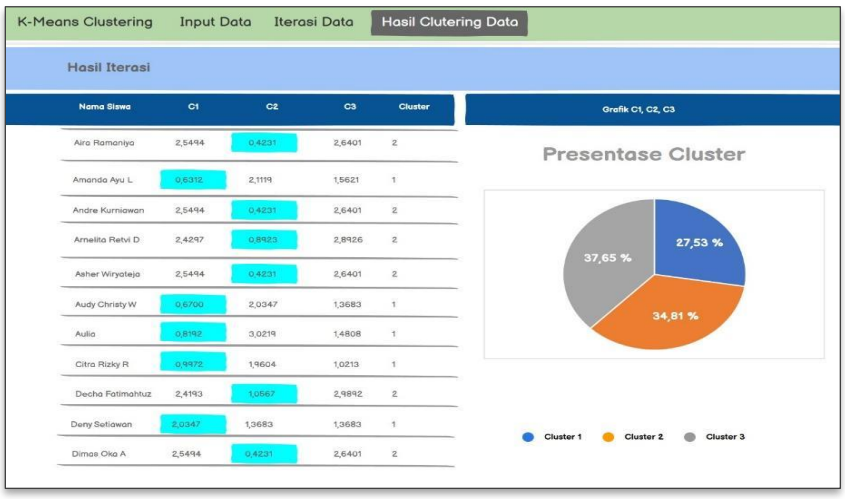

Figure 14. Display of Iteration 5 Results

IV. CLUSTERING RESULT AND COMPARISON

Pada pembahasan ini menjelaskan hasil dari pengolahan data dengan dataset yang sudah diolah dan menghasilkan informasi yang berfungsi sebagai acuan untuk melakukan proses selanjutnya yaitu seleksi siswa program magang industri berdasarkan tiga dataset nilai mata pelajaran, absensi, dan ekstrakurikuler menggunakan metode K-Means Clustering.

Table 13 Reult K-means Clustering

No	Cluster	Class	Count
1	1	10	85
2	2		92
3	3		95
4	1	11	68
5	2		92
6	3		92
7	1	12	51
8	2		74
9	3		92
Total			**741**

Setelah menemukan hasil cluster maksimal, langkah selanjutnya adalah melakukan analisa terhadap setiap cluster untuk menemukan karakteristik siswa setiap cluster yang dapat dilihat secara visualisasi. Analisa ini dibagi berdasarkan nilai mata pelajaran, absensi, dan ekstrakurikuler. Beberapa analisa tersebut diharapkan membentuk karakteristik dari setiap siswa dengan detail

Row No.	id	cluster	mp	absen	ekstra
1	1	cluster_1	2	4	3
2	2	cluster_1	3	4	3
3	3	cluster_1	2	3	3
4	4	cluster_1	2	4	3
5	5	cluster_1	2	4	3
6	6	cluster_1	2	4	3
7	7	cluster_1	2	4	3
8	8	cluster_1	2	4	2
9	9	cluster_0	2	3	2
10	10	cluster_2	2	4	1
11	11	cluster_1	2	4	3
12	12	cluster_2	2	4	1

Figure 15. Clusters of students in grades 10, 11 and 12

Dari hasil seleksi siswa program magang industri, clustering k-means menggunakan Rapidminer Studio 10.2. Dataset yang diolah dibagi menjadi 3 Cluster yaitu dengan menentukan nilai k=3. Pembagian dari 3 Cluster menghasilkan cluster model sebagaimana disajikan pada Gambar 15 dan centroid cluster disajikan pada Gambar 16.

.

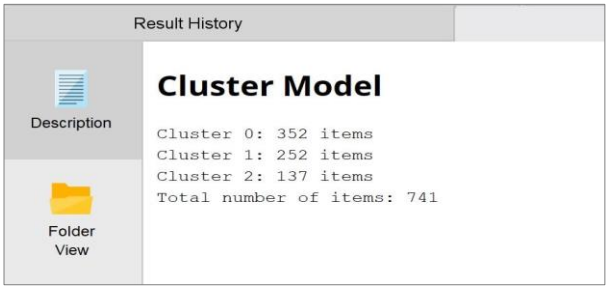

Figure 16. Cluster Grouping Model

Attribute	cluster_0	cluster_1	cluster_2
N. mata pelajaran	1.460	3.857	1.569
N. absensi	3.219	3.194	3.175
N. ekstrakurikuler	3	2.675	1.752

Figure 17. Centroid of Clusters C1, C2, and C3

Dari hasil perhitungan dan evaluasi K-Means Clustering secara konvensional menggunakan microsoft excel dan perhitungan menggunakan rapidminer studio 10.2 dapat dilihat pada tabel 14.

Table 14. Clustering Analysis Results

Analisa	C1	C2	C3	TOTAL	EVALUASI DBI
Microsoft Excel	207	257	277	741	0,729
Rapidminer	352	252	137	741	1,102
Selisih	145	5	140	0	0,373

Tahapan Penggunaan hasil clustering dengan implementasi yang dibuat adalah sebagai berikut:

1. Dataset diinput kedalam aplikasi

Dataset berupa data siswa dengan kriteria penilaian mata pelajaran, absensi kehadiran dan ekstrakurikuler. Input bisa dilakukan per-angkatan kelas 10, 11, dan 12.

2. Pemilihan pusat cluster

Pemilihan pusat cluster yang dalam hal kasus ini adalah terdiri dari 3 cluster. Pemilihan pusat cluster dipilih secara acak, sesuai dataset yang telah diinputkan diproses nomor 1.

3. Proses Iterasi

Setelah kegiatan nomor 2 dilakukan, maka hasil iterasi akan tersaji secara otomatis. Dan terlihat dataset dengan pengelompokan cluster 1, 2 dan 3.

4. Hasil clustering

Hasil clustering terdiri dalam bentuk bagan pie yang merupakan hasil persentase dari perhitungan jarak cluster dan hasil iterasi.

4.1 Visualisasi Penyebaran Data Cluster

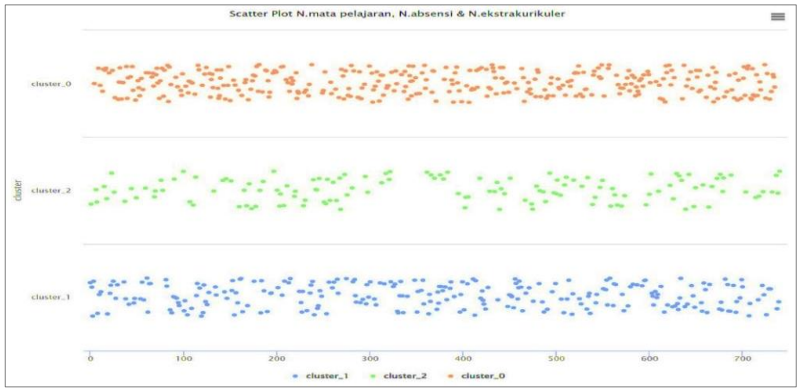

Figure 18. Scatter Plot of N. Subjects, N. Absences, and N. Extracurriculars

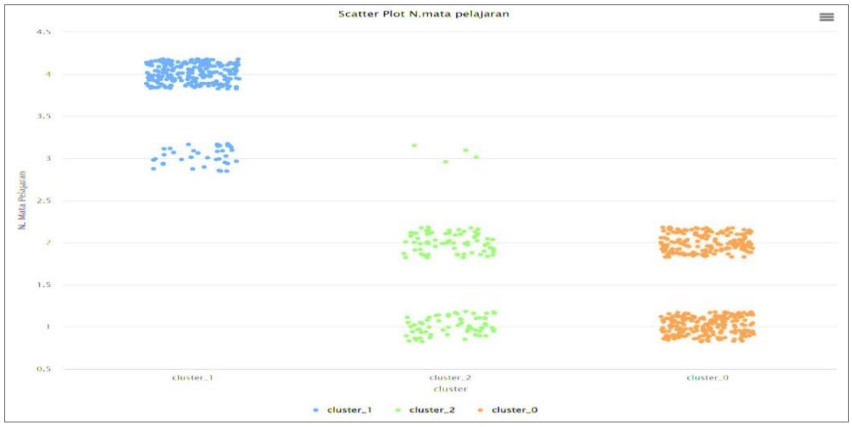

Figure 19. Scatter Plot N. Subjects

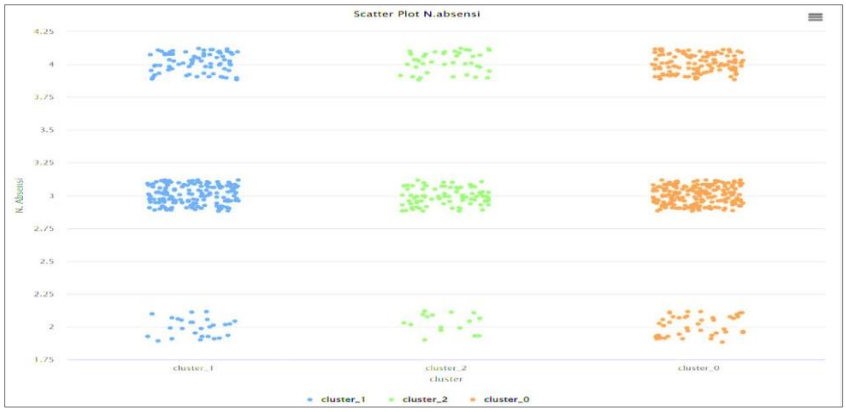

Figure 20. Scatter Plot N. Absence

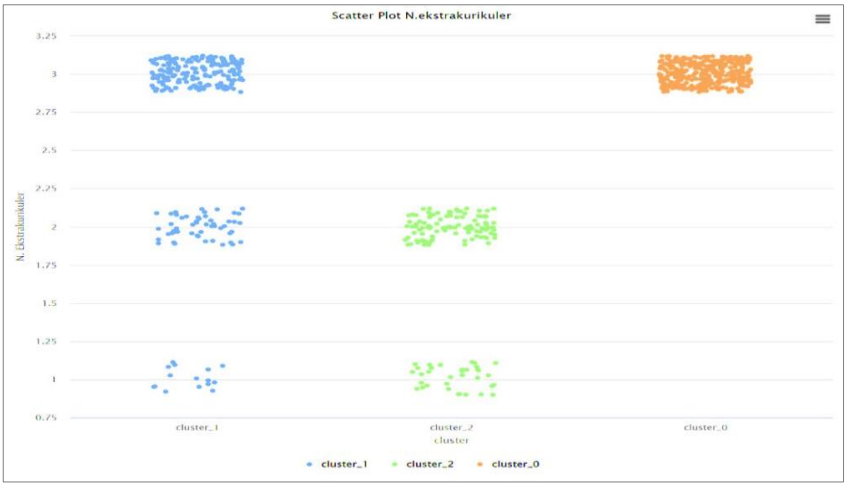

Figure 21. Scatter Plot N. Extracurricular

V. CONCLUSION

1. K-means clustering merupakan metode klasterisasi berdasarkan persamaan karakteristik, dan merupakan metode yang sangat berguna karena mampu mentranslasi ukuran kuantitatif. Berdasarkan hasil pengelompokan data menggunakan algoritma k-means clustering, didapatkan hasil clustering hingga iterasi ke-5, dimana titik pusat tidak lagi berubah dan tidak ada data yang berpindah antar cluster. Penelitian ini menggunakan 741 data uji.
2. Implementasi algoritma k-means clustering ke dalam sistem informasi klasterisasi memberikan hasil klasifikasi pengelompokan data yang efektif dan proses setiap iterasi perputaran jarak centroid, penentuan titik cluster dibentuk, data siswa sebagai acuan objek lebih menghemat waktu melakukan klasterisasi seleksi siswa program magang industri.
3. Proses seleksi siswa program magang industri di SMK Tunas Pemuda Tangerang menggunakan metode Algoritma *k-means clustering* dengan penentuan cluster secara random, menggunakan aplikasi Rapidminer 10.2. Adapun aplikasi tersebut dapat menyeleksi 741 peserta didik yang terbagi ke dalam kelas 10, 11, dan 12 beserta masing-masing memiliki grade standar industri. Hal tersebut dibuktikan dengan perhitungan peserta didik yang menghasilkan cluster yang sama dengan nilai preferensi yang sama dengan perhitungan manual di Microsoft Excel serta menghasilkan nilai akurasi kinerja sistem sebesar 100%.

REFERENCE

[1] Wati, A. R. Z., & Trihantoyo, S. (2020). Strategi pengelolaan kelas unggulan dalam meningkatkan prestasi belajar siswa. JDMP (Jurnal Dinamika Manajemen Pendidikan), 5(1), 46-57.

[2] Antonenko, P. D., Toy, S., & Niederhauser, D. S. (2012). Using cluster analysis for data mining in educational technology research. Educational Technology Research and Development, 60(3), 383-398.

[3] Priyatman, H., Sajid, F., & Haldivany, D. (2019). Klasterisasi Menggunakan Algoritma K-Means Clustering untuk Memprediksi Waktu Kelulusan Mahasiswa. JEPIN (Jurnal Edukasi dan Penelitian Informatika), 5(1), 62-66.

[4] Yang, M. S., & Sinaga, K. P. (2019). A feature-reduction multi-view k-means clustering algorithm. IEEE Access, 7, 114472-114486.

[5] Murai, N., Saito, N., Nii, S., Nishikawa, Y., Suzuki, A., Kodama, E., ... & Nagasaka, S. (2022). Diabetic family history in young Japanese persons with normal glucose tolerance associates with k-means clustering of glucose response to oral glucose load, insulinogenic index and Matsuda index. Metabolism Open, 15, 100196.

[6] Majhi, S. K., & Biswal, S. (2018). Optimal cluster analysis using hybrid K-Means and Ant Lion Optimizer. Karbala International Journal of Modern Science, 4(4), 347-360.

[7] Feng, G., Fan, M., & Chen, Y. (2022). Analysis and Prediction of Students' Academic Performance Based on Educational Data Mining. IEEE Access, 10, 19558-19571.

[8] Liao, Q. Z., Xue, L., Lei, G., Liu, X., Sun, S. Y., & Patil, S. (2022). Statistical prediction of waterflooding performance by K-means clustering and empirical modeling. Petroleum Science

[9] Mughnyanti, M., Efendi, S., & Zarlis, M. (2020). Analysis of determining centroid clustering x-means algorithm with davies-bouldin index evaluation. In IOP Conference Series: Materials Science and Engineering (Vol. 725, No. 1, p. 012128). IOP Publishing

[10] Faid, M., Jasri, M., & Rahmawati, T. (2019). Perbandingan Kinerja Tool Data Mining Weka dan Rapidminer Dalam Algoritma Klasifikasi. Teknika, 8(1), 11-16.

[11] Nirsal, N., Rusmala, R., & Syafriadi, S. (2020). Desain Dan Implementasi Sistem Pembelajaran Berbasis

E-Learning Pada Sekolah Menengah Pertama Negeri 1 Pakue Tengah. d'ComPutarE: Jurnal Ilmiah Information Technology, 10(1), 30-37.

[12] Madyatmadja, E. D., Jordan, S. I., & Andry, J. F. (2021). Big Data Analysis Using Rapidminer Studio To Predict Suicide Rate In Several Countries. ICIC Express Letters, Part B: Applications, 12(8).

[13] Marzukhi, S., Awang, N., Alsagoff, S. N., & Mohamed, H. (2021, August). RapidMiner and Machine Learning Techniques for Classifying Aircraft Data. In Journal of Physics: Conference Series (Vol.1997, No. 1, p. 012012). IOP Publishing.