Klasterisasi Pola Kehadiran Pegawai Institut Teknologi Pagar Alam Menggunakan Algoritma *K-Means*

Tari Kristianda*, Ferry Putrawansyah, Febriansyah*****
* Program Studi Teknik Informatika, Institut Teknologi Pagar Alam (ITPA)
** Program Studi Teknik Informatika, Institut Teknologi Pagar Alam
***Prodi Teknik Informatika, Institut Teknologi Pagar Alam

Correspondence Author :tarikristianda92@gmail.com

Abstract
Tujuan dari penelitian ini adalah untuk mengetahui klasterisasi pola kehadiran pegawai menggunakan Metode *Clustering* dengan algoritma *K-Means* di Institut Teknologi Pagar Alam. Penelitian ini dilatar belakangi dengan proses pengelolaan data kehadiran pegawai yang dilakukan masih hanya sebatas pengarsipan saja, data disimpan dalam bentuk file microsoft *excel*. Dari proses pengumpulan data tersebut belum adanya pengelolaan yang lebih lanjut sehingga dapat menghambat pemantauan kedisiplinan pegawai dan ketaatan dalam hal ketepatan waktu datang. Sedangkan data tersebut dinilai perlu bagi instansi untuk meningkatkan kinerja dari pegawai, untuk menentukan dan membuat kebijakan baru terhadap kinerja pegawai. Data pegawai diolah menggunakan *Rapid Miner* dan bahasa pemrograman *Python,* metode pengembangan sistem dalam penelitian ini menggunakan metode *Cross Industry Standard Process for Data Mining* (*CRISP-DM)*, dimana tahapan meliputi *Business Understanding*, *Data Understanding, Data Preparation, Modelling, Evaluation* dan *Deployment*. Untuk metode pengujian menggunakan *Elbow Method* pengujian menghitung hasil dari *sum of square error (SEE)* dari tiap nilai K untuk mencari jumlah *cluster* terbaik dengan melihat hasil yang berbentuk siku. Hasil dari penelitian ini yakni pola kehadiran pegawai ditahun 2020 yaitu *Cluster_0* dengan Jumlah 6 Pegawai, *Cluster_1* dengan jumlah 10 pegawai, *Cluster_2* dengan jumlah pegawai 9. Selanjutnya ditahun 2021 diperoleh *cluster_0* sebanyak 6 pegawai, *cluster_1* sebanyak 11 pegawai, dan *cluster_2* sebanyak 8 pegawai. Dan ditahun 2022 diperoleh *cluster_0* sebanyak 15 pegawai, *cluster_1* sebanyak 8 pegawai, dan *cluster_2* sebanyak 2 pegawai. Untuk hasil pengujian *Elbow Method* dengan perhitungan *sum of square error (SEE)* yaitu 1249.721. Kemudian diperoleh 3 *cluster* yang tepat yang berbentuk siku diantaranya *cluster_0* dengan keterangan Tepat Waktu, *cluster 1* dengan keterangan Sedang dan *cluster 2* dengan keterangan Tidak Tepat Waktu. Maka hasil klasterisasi dengan *Rapid Miner* dapat dikatakan *valid* atau sesuai dengan pengujian menggunakan *Elbow Method* yang ada di *Python.*

Kata kunci : *Clustering K-Means; CRISP-DM; Elbow Method; Rapid Miner.*

1. PENDAHULUAN

Pesatnya perkembangan *data mining* tidak lepas dari perkembangan teknologi informasi yang memungkinkan terakumulasinya informasi dalam jumlah besar seiring dengan perkembangan teknologi informasi. Data mining telah mendapatkan begitu besar perhatian dan merupakan salah satu teknik yang dapat dimanfaatkan untuk mengolah data agar hasil pengolahan atau informasi yang didapatkan tepat guna. semakin diakui sebagai alat penting untuk manajemen data seiring dengan bertambahnya jumlah data. *Data Mining* sering disebut *Knowledge Discovery in Database (KDD).* [1].

Data mining adalah proses pencarian informasi atau pola yang menarik di dalam data, atau disebut juga dengan istilah yang digunakan untuk mencari informasi di dalam database dengan menggunakan teknik atau metode tertentu. Penambangan data adalah penerapan kecerdasan buatan, pembelajaran mesin, teknik statistik dan matematika untuk mengidentifikasi dan menemukan informasi berguna untuk informasi yang terkait dengan berbagai database berkapasitas tinggi [2]. Data mining merupakan proses informasi dengan mencari pola dan hubungan yang tersembunyi pada tumpukan data. Data mining disebut juga sebagai *Knowledge In Database (KDD)* yaitu kegiatan pengumpulan pemakaian data lampau untuk menemukan pola atau hubungan terhadap data yang ukurannya besar [3]. Ada beberapa metode yang bisa dipakai dalam data mining salah satunya adalah *Clustering*.

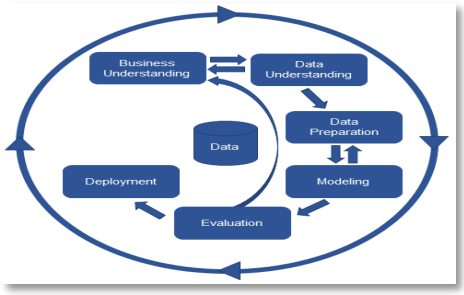
Clustering atau analisis kelompok adalah metode membagi kumpulan data menjadi beberapa kelompok berdasarkan kesamaan yang diberikan. Objek dikelompokkan menjadi satu atau lebih cluster sehingga objek dalam satu cluster sangat mirip satu sama lain. [4]. Cluster dapat dipakai lebih lanjut dalam berbagai aplikasi secara luas seperti klasifikasi, pengolahan gambar, dan pengenalan pola. Salah satu metode untuk melakukan *cluster* adalah *K-Means*. *K-Means* merupakan salah satu algoritma data mining yang dapat digunakan untuk mengelompokkan atau mengelompokkan data. K-Means merupakan metode *partial clustering* yang juga dapat digunakan untuk membagi data menjadi kelompok/cluster. [5].

Berdasarkan hasil wawancara dan observasi pada bagian kepegawaian Institut Teknologi Pagar Alam (Wita Hariani, 2022), saat ini proses pengelolaan data kehadiran pegawai yang dilakukan masih hanya sebatas pengarsipan saja, data disimpan dalam bentuk file *microsoft excel*. Proses absensi yang berjalan saat ini digunakan untuk perhitungan biaya transport, melihat tingkat kehadiran, dan penilaian terhadap struktural atau pegawai. Waktu

masuk pegawai Institut Teknologi Pagar Alam dari jam 08.00 WIB, batas keterlambatan hadir paling telat adalah 08.15 WIB sampai dengan pulang jam 15.00 WIB. Setiap pegawai yang terlambat akan mendapatkan sanksi berupa himbauan atau tegoran langsung. Dari proses pengumpulan data tersebut belum adanya pengelolaan yang lebih lanjut sehingga dapat menghambat pemantauan kedisiplinan pegawai dan ketaatan dalam hal ketepatan waktu datang. Sedangkan data tersebut dinilai perlu bagi instansi untuk meningkatkan kinerja dari pegawai, untuk menentukan dan membuat kebijakan baru terhadap kinerja pegawai. Maka dari itu dibutuhkan sebuah data mining dengan metode *Clustering* menggunakan *Algoritma K-Means* untuk mengetahui pola kehadiran pegawai yang ada di Insitut Teknologi Pagar Alam, dimana setelah mengetahui pola kehadiran dapat dijadikan sebagai pedoman bagi pihak instansi untuk melakukan kebijakan selanjutnya.

2. METODE PENELITIAN

Metodologi penelitian adalah ilmu yang mempelajari metode penelitian dan pengetahuan tentang alat penelitian. Metodologi penelitian melihat konsep teoretis dari berbagai metode dan, oleh karena itu, keuntungan dan kerugian apa yang harus dipertimbangkan ketika memilih metode untuk penulisan ilmiah. Cabang metode penelitian terbagi menjadi dua bidang, yaitu metodologi penelitian kuantitatif dan penelitian kualitatif. (Albi Anggito, 2018). Metode penelitian yang digunakan dalam penelitian yaitu kualitatif karena bersifat deskripsi atau menggambarkan, cenderung menggunakan analisis. Dimana fokus utamanya mencari informasi yang lengkap mengenai penelitian yang dilakukan. Dalam penyusunan penelitian ini, menggunakan tahapan dalam *CRISP-DM*. CRISP-DM *(Cross Industry Standard Process for Data Mining)* suatu standarisasi pemrosesan data mining yang telah dikembangkan dimana data yang ada akan melewati setiap fase terstruktur dan terdefinisi dengan jelas dan efisien. Selain menerapkan suatu model dalam proses penambangan data, pemilihan algoritma sangat mempengaruhi terhadap komparasi kinerja metode data mining. Metodologi ini terdiri dari enam tahapan yaitu *Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, dan Deployment* [6]. Tahapan penelitan yang dilakukan sebagai berikut:



Gambar 1. Model *Crisp-DM*

1. Pemahaman Bisnis *(Business Understanding)*
   Langkah pertama dari topik penelitian ini melibatkan data kehadiran karyawan. Pada titik ini perlu dipahami pentingnya pemanfaatan data pegawai agar dapat digunakan untuk membuat kebijakan yang lebih obyektif sehingga dapat meningkatkan kinerja pegawai. Oleh karena itu, diperlukan strategi sintesis atau sintesa data pegawai.
2. Pemahaman data *(Data Understanding)*
   Peneliti melakukan pemahaman terhadap kebutuhan data terkait dengan kehadiran pegawai di Institut Teknologi Pagar Alam. Data diambil dari bagian kepegawaian, pemahaman data mengacu pada klasterisasi kehadiran dengan atribut seperti PIN, NIP, Jabatan, Departemen, Kantor, tanggal, Jam masuk, dan Jam pulang. Setelah data didapatkan dilakukan eksplorasi data sejumlah 37 record pegawai. Data yang dikumpulkan yaitu data kehadiran pegawai selama 3 Tahun pada tahun 2020 sampai dengan 2022.

Tabel 1. Data Kehadiran 2020-2022

No	Pin	NIP	Nama	Jabatan	Departemen	Kantor	Tanggal	Jam Masuk	Jam Pulang
1	1		Dosen 1				12-01-2020	15:59:19	11:48:15
2	1		Dosen 1				13-01-2021	08:27:03	16:09:30
....	....		.....				.....	.....	.....
15018	1		Dosen 1				16-01-2022	08:10:45	16:22:22

3. Persiapan Data *(Data Preparation)*
   Setelah melakukan fase pemahaman dari data kehadiran pegawai, maka pada fase pengolahan data ini membuat *dataset* final yang akan diterapkan ke dalam pemodelan data. Dalam tahapan ini yaitu membangun *dataset* akhir dari berupa data mentah. Ada beberapa hal yang dilakukan antara lain pembersihan data, pembersihan record, kemudian dilakukan seleksi data, *record* dan atribut. Data tersebut kemudian ditransformasikan menurut kriteria tertentu (transformasi data). Kriteria evaluasi yang digunakan adalah:

Tabel 2. Kriteria yang digunakan

Kriteria	Jam Masuk	Jam Pulang	Nilai
Tepat waktu	07.30 - 08.00	> 15.00	3
Sedang	08.00 - 08.15	13.00 - 15.00	2
Tidak tepat waktu	> 08.15	< 13.00	1

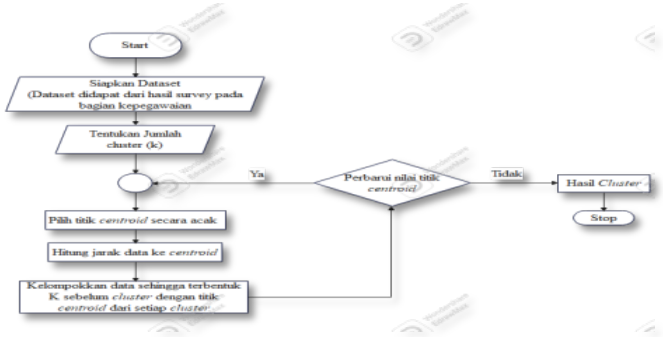
Adapun transformasi data sample kehadiran yang dilakukan selama tahun 2020 sebagai *sample* adalah sebagai berikut :

Tabel 3. Transformasi *sample* kehadiran tahun 2020

PIN	Nama	Januari	Februari	Maret	April	Mei	Juni	Juli	Agustus	September	Oktober	November	Desember
1	Ferry	3	3	2	3	3	3	3	2	3	2	3	2
2	Admin	1	1	1	2	1	2	1	1	1	1	1	1
3	Febri	3	3	3	2	2	3	3	3	3	3	3	3

4. Pemodelan *(Modelling)*

Adapun pemodelan yang dilakukan pada penelitian ini adalah metode *clustering* menggunakan algoritma *k-means.* Berikut langkah-langkah algoritma *K-Means* yang digunakan dalam pemodelan ini :



Gambar 2. Langkah-langkah algoritma *k-means*

a. Tentukan Jumlah *Cluster*

Berdasarkan *dataset* yang disiapkan, digunakan cluster yang dibentuk berdasarkan 3 kelompok pegawai yakni kelompok Tepat Waktu, Sedang, dan Tidak Tepat Waktu pada Institut Teknologi Pagar Alam.

C0 : Tepat Waktu
C1 : Sedang
C2 : Tidak Tepat Waktu

b. Tentukan *centroid* secara acak

Dibutuhkan 3 *centroid* sebagai titik pusat *cluster* yang ditentukan secara acak, karena ada 3 kelompok atau cluster

c. Hitung jarak dari data ke centroid

Untuk menghitung jarak ke centorid, digunakan rumus jarak Euclid:

$$d(x,y) = i\sum_{k=1}^{n}(x_i - y_i)^2$$

Keterangan :
d(x,y) = jarak antara data x ke data y
xi = data testing ke-i
yi = data training ke-i

Pada perhitungan ini adalah menentukan nilai titik tengah cluster awal untuk setiap cluster dari setiap variabel. Nilai awal pusat cluster pada iterasi pertama (perhitungan pertama) diberikan secara acak. Pada iterasi selanjutnya, nilai awal pusat cluster (iterasi pertama sampai titik iterasi normal/maksimal) diperoleh dengan cara merata-ratakan data tiap cluster. Jika nilai pusat klaster awal yang baru sama dengan nilai pusat klaster awal yang baru, proses iterasi berlanjut hingga nilainya sama atau hingga nilai maksimum iterasi yang telah ditentukan sebelumnya. Jika nilai pusat klaster asli yang baru sama dengan pusat klaster asli yang lama, maka proses pengelompokan berhenti.

d. Perbarui titik *Centroid*

Perhitungan titik *centroid* baru dengan berdasarkan pengelompokkan yang pertama untuk melakukan iterasi selanjutnya.

e. Ulangi langkah 3, 4, 5 sampai nilai dari *titik centroid* tidak lagi berubah.

Perhitungan dilakukan sampai jumlah data ditentukan. Setelah menghitung jarak data dari pusat atau center, langkah selanjutnya adalah mengelompokkan data berdasarkan jarak terdekat dari pusat. Anda kemudian dapat menjalankan iterasi berikutnya untuk membandingkan hasilnya. Jika titik tengahnya berubah, ulangi langkah tersebut hingga titik tengahnya tidak berubah lagi.

5. *Evaluation*

Pada tahap ini lebih menitikberatkan pada model yang dihasilkan sesuai dengan standar kluster *K-Means,* dan pada tahap pertama tidak ada yang dihilangkan hingga tahap pemodelan selesai dengan pengujian dengan metode *elbow*.

6. *Deployment*

Pada tahap akhir ini, CRISP-DM meliputi diseminasi pengetahuan atau informasi yang diperoleh, yang selanjutnya akan diimplementasikan dalam bentuk laporan dan analisis terhadap setiap cluster yang dihasilkan sehingga Institut Teknologi Pagar Alam dapat mudah untuk memahaminya.

3. HASIL DAN ANALISA

3.1 Hasil

Berdasarkan penelitian tersebut, kami membuat model clustering pola kehadiran karyawan dengan menggunakan algoritma k-means modeling. Model ini dapat dijadikan usulan klaster yang ideal untuk mengkaji pola kehadiran pegawai selama 3 tahun, tepatnya pada tahun 2020 hingga tahun 2022 dengan jumlah pegawai sebanyak 25 orang. Hasil dari pola tersebut ada 3 *cluster*/kelompok tersebut yaitu kelompok C0 yang merupakan kelompok pegawai yang hadir tepat waktu, C1 yang merupakan kelompok pegawai dengan kehadiran sedang dan C2 yang merupakan kelompok pegawai yang sering hadir tidak tepat waktu. Hasil *Cluster* yang pertama di tahun 2020 yaitu cluster_0 yang berjumlah 6 pegawai dengan tingkat kehadiran Tepat Waktu, *cluster_1* berjumlah 10 pegawai dengan tingkat kehadiran Sedang, dan *cluster_2* berjumlah 9 pegawai dengan tingkat kehadiran Tidak Tepat Waktu. Kemudian untuk tahun 2021 cluster_0 yang berjumlah 6 pegawai, cluster_1 berjumlah 11 pegawai, dan *cluster_2* berjumlah 8 pegawai. Kemudian untuk tahun 2022 *cluster_0* berjumlah 15 pegawai yang Tepat Waktu, *cluster_1* berjumlah 8 pegawai dengan tingkat kehadiran sedang, dan *cluster_2* berjumlah 2 pegawai dengan tingkat kehadiran Tidak Tepat Waktu. Kemudian hasil dari metode pengujian yang diaplikasikan kedalam bahasa pemrograman *Python* menggunakan *Elbow Method* dengan menghitung hasil *Sum of square error (SEE)* adalah 1249.721. Maka diperoleh nilai K untuk mencari jumlah *cluster* terbaik dengan melihat hasil yang berbentuk siku yaitu K=3 dengan keterangan *cluster_0* yakni memiliki tingkat pola kehadiran Tepat Waktu, cluster_1 dengan keterangan tingkat pola kehadiran sedang dan cluster_2 dengan keterangan tingkat pola kehadiran Tidak Tepat Waktu. Sehingga dapat dikatakan *valid* atau sesuai dengan hasil *clastering k-means* menggunakan *Rapid Miner*.

3.2 Pembahasan

Menggunakan Tahapan *Crisp-DM* diantaranya :

a. Pemahaman Bisnis

Proses pengelolaan data kehadiran untuk mengetahui pegawai mana yang sering terlambat masih dilakukan secara manual, khususnya visualisasi dan perhitungan di excel. Dari proses pengelompokan ini, tidak ada cara manajemen lain yang dapat menghambat pengendalian disiplin pegawai. Oleh karena itu, peneliti akan mengklasterisasi pegawai di ITPA dengan menggunakan metode k-means untuk mengidentifikasi pola kehadiran pegawai sehingga dapat digunakan oleh instansi untuk meningkatkan kinerja dari pegwai.

b. Pemahaman Data

Pada fase pemahaman data ini, data yang telah didapat dari bagian kepegawaian Institut Teknologi Pagar Alam. Data yang diambil yaitu data kehadiran pegawai selama 3 tahun pada tahun 2020 sampai dengan tahun 2022 ada 15018 *reccord* dengan 10 atribut yaitu PIN, NIP, Nama, Jabatan, Departemen, Kantor, Tanggal, Jam Masuk, Jam Pulang, dan Jam Lembur, kategori data yang diterima dalam bentuk *excel* dan data yang didapat perlu di *cleaning* dan akan dilakukan pemilihan data dengan atributnya.

c. Pengolahan Data

Pada tahap ini diperoleh 10 atribut dari bagian Sumber Daya Manusia Institut Teknologi Pagar Alam dengan catatan data kehadiran sebanyak 15.018 orang. Seleksi kemudian dilakukan di Excel dengan memfilter. Setelah menyaring data, peneliti menggunakan lima atribut pada tahap pemilihan data, antara lain PIN, nama, waktu masuk, dan waktu keluar, dengan jumlah karyawan yang dikelompokkan masing-masing 25 orang per tahun. Dan hasil *Cleaning* menggunakan *excel* dari

15.018 data menjadi 10.091 yang akan dijadikan *datasheet.* Sebelumnya peneliti sudah melakukan analisa kehadiran perbulan dari setiap pegawai. Kemudian hasil kehadiran perbulan didapat rata-rata kehadiran yang akan di jadikan *datasheet* pertahun. Karena total seluruh data berjumlah 10.091, sedangkan aplikasi *Rapid Miner* hanya membaca sebanyak 10.000 data maka dari itu peneliti melakukan proses *clustering* berdasarkan tahun. Dan agar bisa melihat perbandingan dari kehadiran pegawai disetiap tahunnya.

Tabel 4. *Datasheet* kehadiran tahun 2020

PIN	Nama	Januari	Februari	Maret	April	Mei	Juni	Juli	Agustus	September	Oktober	November	Desember
1	Ferry	3	3	2	3	3	3	3	2	3	2	3	2
2	Admin	1	1	1	2	1	2	1	1	1	1	1	1
3	Febri	3	3	3	2	2	3	3	3	3	3	3	3

d. Pemodelan

Berikut ini adalah langkah-langkah dalam mengetahui pola kehadiran pegawai menggunakan algoritma *KMeans Clustering* pada aplikasi *Rapid Miner* dan *Python.*

a. Tentukan Jumlah *Cluster*

Setelah menyiapkan *dataset* maka langkah selanjutnya adalah menentukan jumlah *cluster d*isini peneliti menggunakan 3 *cluster.* Untuk menentukan jumlah *cluster* dilakukan beberapa percobaan yang dapat dilihat pada fokus terkecil.

b. Tentukan Titik *Centorid*

Pertama klasterisasi kehadiran untuk tahun 2020 Setelah dilakukan beberapa percobaan, percobaan dengan 3 *cluster* yang paling tepat. Dengan pengukuran *performance vector* rata-rata dalam *centroid distance* yang didapat dengan nilai 3.013, kemudian rata-rata *cluster_0* dengan nilai 2.528, rata-rata *cluster_1* dengan nilai 3.060, dan rata-rata dalam *centroid cluster_2* dengan nilai 3.284 dan nilai *Device Bouldi Index* adalah 1.549.

Attribute	cluster_0	cluster_1	cluster_2
Januari	3	2.500	1.778
Februari	3	1.700	1.556
Maret	2.667	1.500	1.778
April	2.167	1.900	1.667
Mei	2.333	1.800	1.444
Juni	2.333	2	1.444
Juli	2.667	2.400	1.667
Agustus	2	1.900	1.778
September	2.667	2.200	1.333
Oktober	2.667	2.800	1.556
November	2.500	2.900	1.667
Desember	2.167	2.200	1.667

Gambar 3. Titik *Centroid* kehadiran tahun 2020

Percobaan 3 *cluster* itu C0, C1, dan C2 dengn atribut PIN, Nama, kehadiran selama tahun 2021. Dengan pengukuran *performance vector* rata-rata dalam *centroid distance* yang didapat dengan nilai 3.344, kemudian rata-rata *cluster_0* dengan nilai 2.328, rata-rata *cluster_1* dengan nilai 3.781, dan rata-rata dalam *centroid cluster_2* dengan nilai 3.729 dan nilai *Device Bouldi Index* adalah 1.595.

Attribute	cluster_0	cluster_1	cluster_2
Januari	2.667	2.182	2
Februari	2.333	1.455	1.250
Maret	2.500	1.727	1
April	3	2	1.250
Mei	2.333	2.182	1.125
Juni	2.500	2.182	1.625
Juli	3	2	1.625
Agustus	2.833	1.909	1.500
September	2.833	1.727	1.500
Oktober	3	2.636	1.750
November	3	2.273	1.750
Desember	3	2.273	2

Gambar 4. Titik *Centroid* Kehadiran 2021

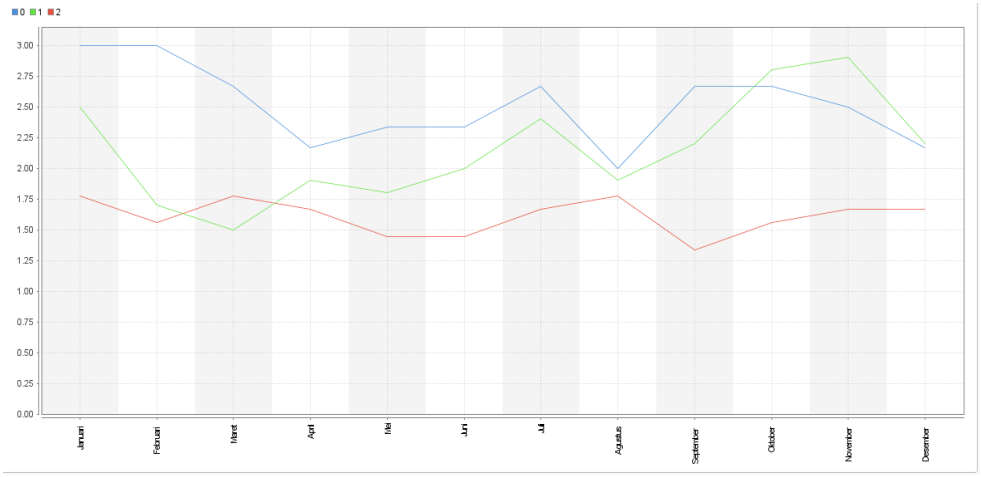
Percobaan 3 *cluster* itu C0, C1, dan C2 dengn atribut PIN, Nama, dan kehadiran selama 2022. Dengan pengukuran *performance vector* rata-rata dalam *centroid distance* yang didapat dengan nilai 1.856, kemudian rata-rata *cluster_0* dengan 1.867, rata-rata *cluster_1* dengan nilai 1.500, dan rata-rata dalam *centroid cluster_2* dengan nilai 1.910 dan nilai *Device Bouldi Index* adalah 1.042.

Attribute	cluster_0	cluster_1	cluster_2
Januari	2.933	2	1.500
Februari	2.933	2	1.500
Maret	2.733	1.875	1.500
April	2.867	2	2.500
Mei	2.800	1.875	2
Juni	2.667	2.250	2
Juli	2.867	2.250	2
Agustus	2.733	2	2
September	2.733	2.250	2.500
Oktober	2.933	2.375	1.500
November	2.800	2.375	1

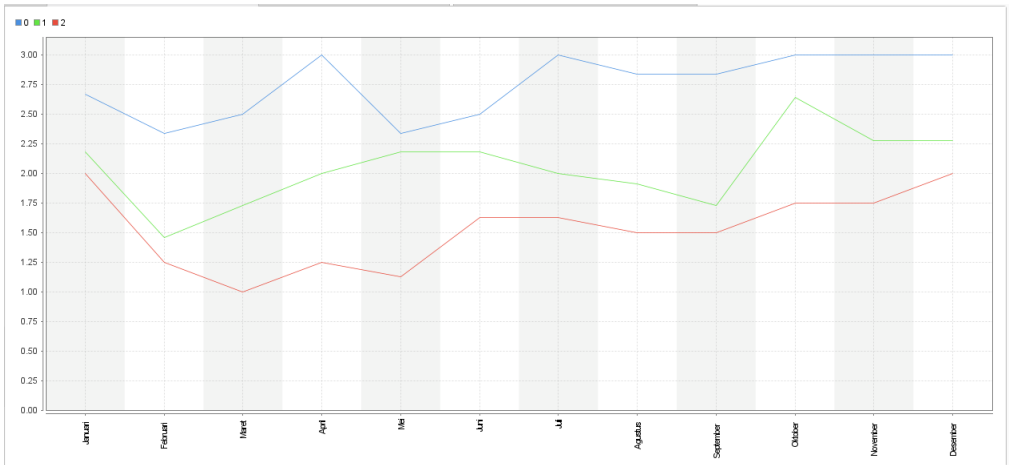
Gambar 5. Titik *Centroid* Kehadiran 2022

c. Menghitung Jarak dengan *Euclidean Distance*

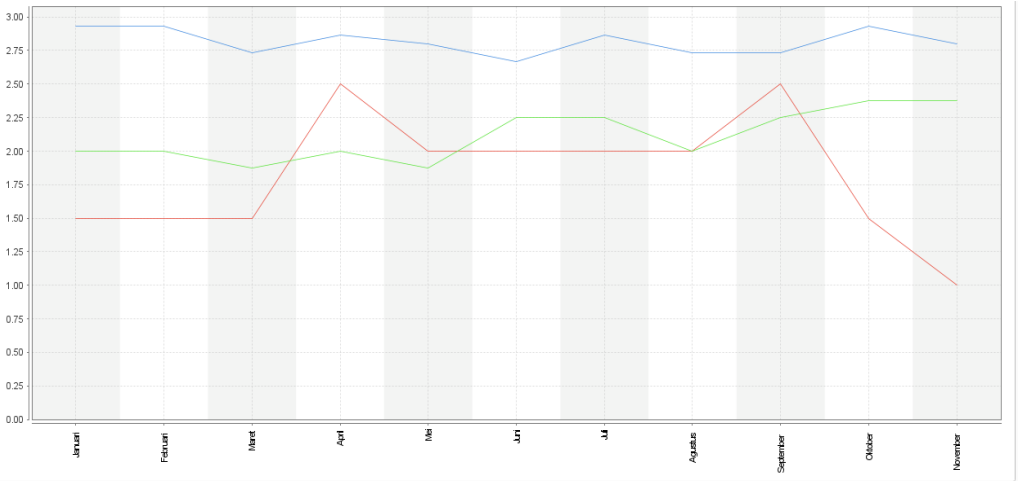
Setelah mengimpor *dataset* dan algoritma *k-means* ke dalam Rapid Miner, proses data dengan mengklik Run untuk menampilkan jarak data ke *centroid*.



Gambar 6. Jarak data ke pusat *centroid* Tahun 2020



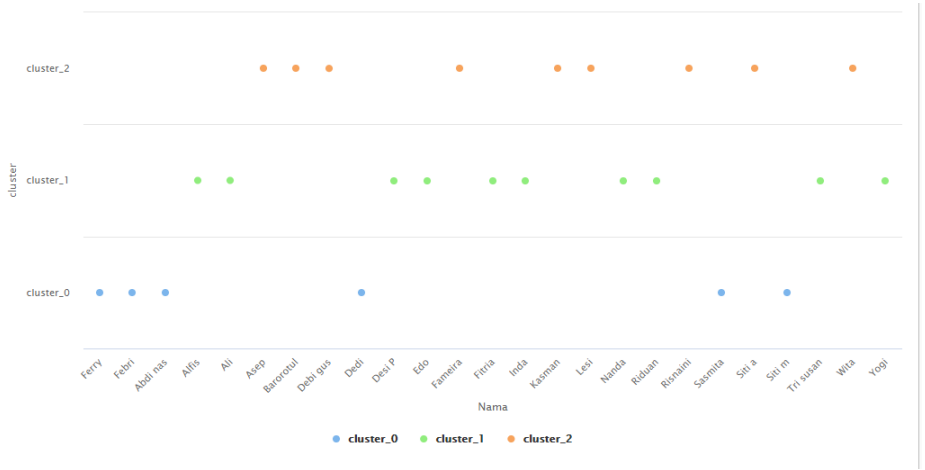
Gambar 7. Jarak data ke *centroid* Tahun 2021



Gambar 8. Jarak data ke *centroid* Tahun 2022

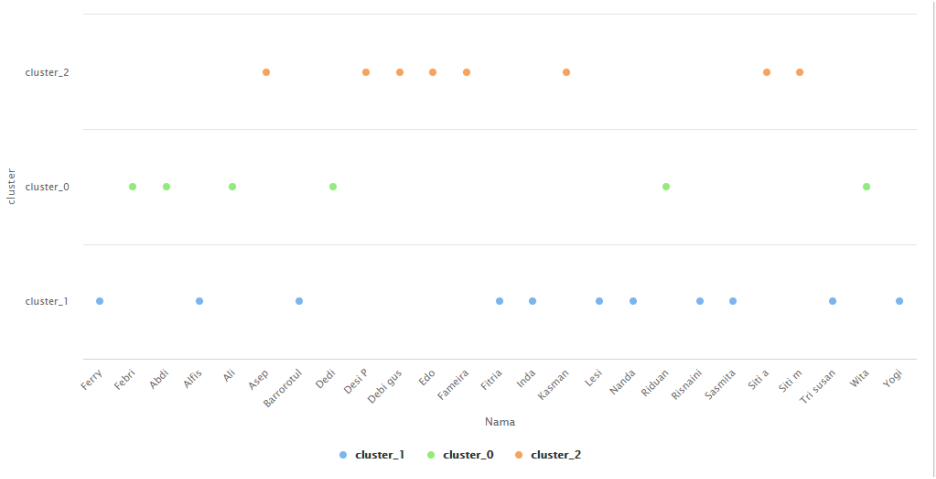
d. Perbarui nilai titik *centroid*
e. Ulangi langkah 3 sampai 5 sampai nilai titik centroid tidak berubah

Jika titik pusat centroid berubah maka iterasi diulangi, tetapi jika tidak berubah maka iterasi dihentikan dan diperoleh hasil dari masing-masing kelompok.



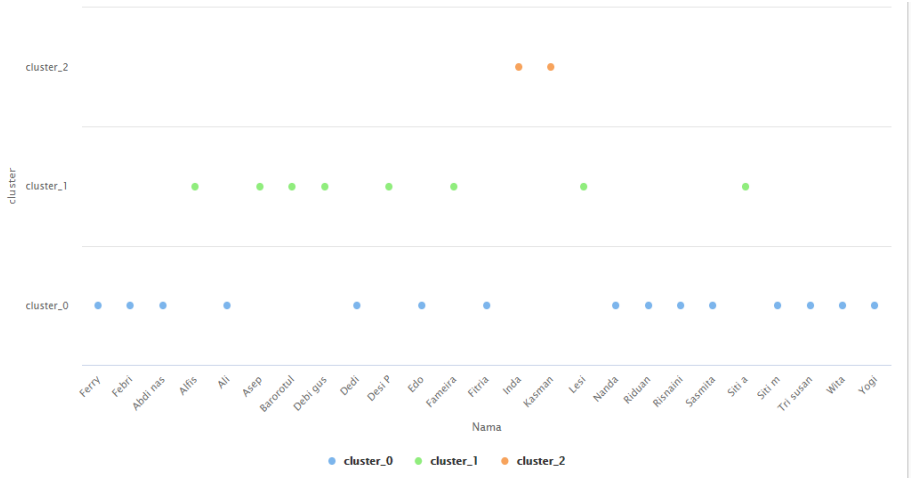
Gambar 9. *Clustering* kehadiran tahun 2020

Gambar diatas merupakan hasil *clustering* pola kehadiran pegawai Institut Teknologi Pagar Alam di tahun 2020 yaitu *cluster_0* dengan kehadiran tepat waktu sebanyak 6 pegawai, *cluster_1* dengan kehadiran sedang sebanyak 10 pegawai dan *cluster_2* dengan kehadiran tidak tepat waktu sebanyak 9 pegawai.



Gambar 10. Hasil *clustering* kehadiran tahun 2021

Kemudian hasil *clustering* pola kehadiran pegawai Institut Teknologi Pagar Alam di tahun 2021 yaitu *cluster_0* dengan kehadiran tepat waktu sebanyak 6 pegawai, *cluster_1* dengan kehadiran sedang sebanyak 11 pegawai dan *cluster_2* dengan kehadiran tidak tepat waktu sebanyak 8 pegawai.



Gambar 11. Hasil *clustering* kehadiran tahun 2022

Dan terakhir hasil *clustering* pola kehadiran pegawai Institut Teknologi Pagar Alam di tahun 2022 yaitu *cluster_0* dengan kehadiran tepat waktu sebanyak 15 pegawai, *cluster_1* dengan kehadiran sedang sebanyak 8 pegawai dan *cluster_2* dengan kehadiran tidak tepat waktu sebanyak 2 pegawai.

Berdasarkan perhitungan Rapid Miner, diperoleh model yang kemudian diimplementasikan dengan Python. Berdasarkan hasil perhitungan jarak, model yang digunakan untuk mengelompokkan data adalah :

a. Jika C0 <; C1 dan C0 <; C2 maka *Cluster_0* dengan keterangan Tepat Waktu
b. Jika C1 <; C0 dan C1 <; C2 kemudian *Cluster_1* dengan keterangan Sedang
c. Jika C1 <; C0 dan C2 < C1 lalu *Cluster_2* dengan keterangan Tidak Tepat Waktu

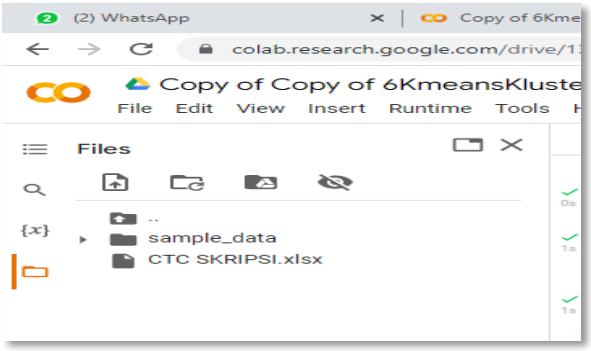
Dari data kehadiran yang diolah dengan *Rapid Miner* yang digunakan pada sistem menggunakan metode *Clustering K-Means* maka diperoleh pola kehadiran pegawai ditahun 2020 yaitu *Cluster_0* memiliki tingkat kehadiran Tepat Waktu dengan Jumlah 6 Pegawai, *Cluster_1* memiliki tingkat kehadiran Sedang dengan jumlah 10 pegawai, *Cluster_2* memiliki tingkat kehadiran Tidak Tepat Waktu dengan jumlah pegawai 9. Selanjutnya ditahun 2021 diperoleh *cluster_0* sebanyak 6 pegawai dengan tingkat kehadiran Tepat Waktu, *cluster_1* sebanyak 11 pegawai dengan tingkat kehadiran sedang, dan *cluster_2* sebanyak 8 pegawai dengan keterangan Tidak Tepat Waktu. Dan ditahun 2022 diperoleh *cluster_0* sebanyak 15 pegawai dengan tingkat kehadiran Tepat Waktu, *cluster_1* sebanyak 8 pegawai dengan tingkat kehadiran sedang, dan *cluster_2* sebanyak 2 pegawai dengan keterangan Tidak Tepat Waktu. Setelah dilakukan proses *Clustering* maka dapat diketahui *Cluster* Pola kehadiran pegawai berdasarkan tingkat kehadirannya.

e. Evaluation

Setelah hasil *cluster* didapatkan dari kehadiran pegawai selama 3 tahun yang sudah dimodelkan sesuai dengan standar *K-Means Clustering*. Model algoritma *k-means clustering* dapat diterapkan pada bahasa pemrograman *Python* menggunakan *Google Colab.* Berikut penerapan *k-means clustering* :

a. Halaman *Import File Google Colab*

Pada halaman ini terlebih dahulu memasukkan file data kehadiran selama 3 tahun data dengan format *excel.*



Gambar 12.*Gambar Menu* Import File

b. Memanggil *datasheet* dari *Library Python*

Setelah file data kehadiran selama 3 tahun di *import*-kan ke dalam *library*, selanjutnya melakukan pemanggilan data. Berikut *coding* untuk memanggil data.

```
[1]  # https://www.reneshbedre.com/blog/kmeans-clustering-python.html

[2]  from sklearn.datasets import make_blobs
     import pandas as pd

[3]  df = pd.read_excel('CTC SKRIPSI.xlsx') #memanggil dataset
```

Gambar 13. *Coding Import*

```
] df.head()
```

	PIN	Januari	Februari	Maret	April	Mei	Juni	Juli	Agustus	September	Oktober	November	Desember	Tahun
0	1	3	3	2	2	3	2	2	2	2	2	2	2	2020
1	3	3	3	3	2	2	3	3	3	3	3	3	3	2020
2	34	3	3	3	3	3	3	3	2	3	3	3	2	2020
3	13	2	2	2	2	2	1	2	2	2	3	2	2	2020
4	33	3	1	2	3	3	2	3	2	2	3	3	3	2020

Gambar 14. *Datasheet*

c. Menentukan nilai *K-Means Clustering* ke *dataset*

Berikut program untuk menjalankan *K-Means Clustering*, disini peneliti menggunakan 3 *cluster.*

```
from sklearn.cluster import KMeans #tentukan dan mengkonfigurasi fungsi k-means
kmeans = KMeans(n_clusters=3, init='k-means++', random_state=0).fit(df)
```

Gambar 15. *K-Means Clustering*

d. Hasil perhitungan dari pusat *Cluster* atau titik *centroid*

Berikut perhitungan untuk nilai titik *centroid :*

```
[56]  kmeans.n_iter_
0s
        4
```

Gambar 16. Titik Pusat *Cluster*

```
[70] kmeans.cluster_centers_

array([[3.06666667e+01, 2.22222222e+00, 2.00000000e+00, 2.05555556e+00,
        2.27777778e+00, 2.11111111e+00, 2.16666667e+00, 2.38888889e+00,
        2.16666667e+00, 2.16666667e+00, 2.22222222e+00, 2.16666667e+00,
        2.27777778e+00, 2.02100000e+03],
       [7.09090909e+00, 2.60606061e+00, 2.18181818e+00, 2.09090909e+00,
        2.15151515e+00, 2.18181818e+00, 2.33333333e+00, 2.42424242e+00,
        2.21212121e+00, 2.30303030e+00, 2.75757576e+00, 2.54545455e+00,
        2.36363636e+00, 2.02100000e+03],
       [2.05000000e+01, 2.16666667e+00, 1.83333333e+00, 1.75000000e+00,
        2.04166667e+00, 1.79166667e+00, 1.87500000e+00, 2.08333333e+00,
        1.91666667e+00, 1.95833333e+00, 2.25000000e+00, 2.33333333e+00,
        1.95833333e+00, 2.02100000e+03]])
```

Gambar 17. Perhitungan pusat *cluster*

e. Hasil perhitungan object *cluster*

Dari hasil data kehadiran selama 3 tahun diperoleh hasil perhitungan *cluster* terbanyak yaitu *cluster_1* dengan tingkat kehadiran sedang.

```
from collections import Counter
Counter(kmeans.labels_)

Counter({1: 33, 0: 18, 2: 24})
```

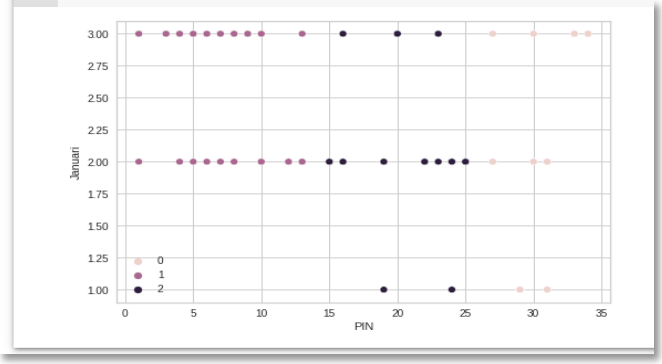
Gambar 18. Jumlah *Cluster*

f. Menampilkan hasil *cluster* Kehadiran

Dibawah ini merupakan program untuk menampilkan hasil *cluster* disetiap bulan dalam waktu 3 tahun.

```
import seaborn as sns
import matplotlib.pyplot as plt
sns.scatterplot(data=df, x="PIN", y="Januari", hue=kmeans.labels_)
plt.show()
```

Gambar 19. Coding untuk menampilkan hasil *cluster*



Gambar 20. Hasil *Clustering*

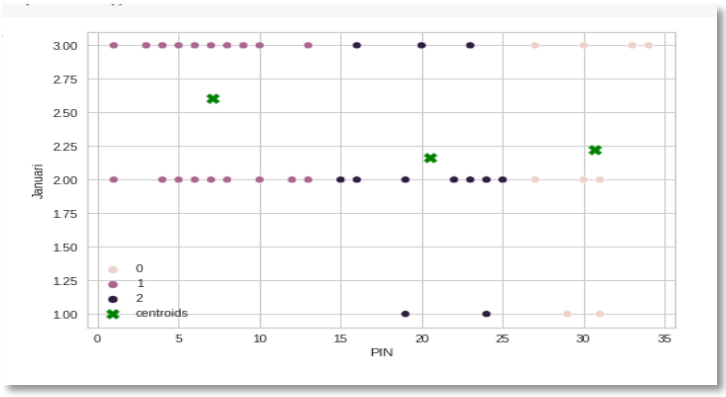
g. Menampilkan Titik *centroid*

Selanjutnya kita bisa menampilkan hasil dari titik *centroid* kehadiran pegawai. Berikut *coding* program unuk menampilkan hasil *centroid.*

```
sns.scatterplot(data=df, x="PIN", y="Januari", hue=kmeans.labels_)
plt.scatter(kmeans.cluster_centers_[:,0], kmeans.cluster_centers_[:,1],
            marker="X", c="green", s=100, label="centroids")
plt.legend()
plt.show()
```

Gambar 21. Program Untuk menampilkan *centroid*

Dari program diatas didapatkan hasil seperti gambar dibawah ini :



Gambar 22. Hasil titik *Centorid*

Langkah selanjutnya adalah evaluasi menggunakan metode *Elbow* yang diprogram dengan Python melalui *Google Collaboration*. Metode siku merupakan suatu metode untuk menentukan jumlah cluster yang tepat atau optimal berdasarkan persentase hasil perbandingan antara jumlah cluster yang membentuk siku pada suatu titik. Dengan menghitung hasil dari *Sum of Square Error (SSE) :*

Dengan rumus :

$$SSE = \sum_{k=1}^{k} \sum_{xi} [Xi - Ck]^2$$

Beberapa tahapan metode *elbow* yaitu inisialisasi nilai awal k, naikan nilai k, kemudian hitung hasil *sum of square error (SSE)*, melihat hasil *SSE*, tetapkan nilai k yang berbentuk siku.

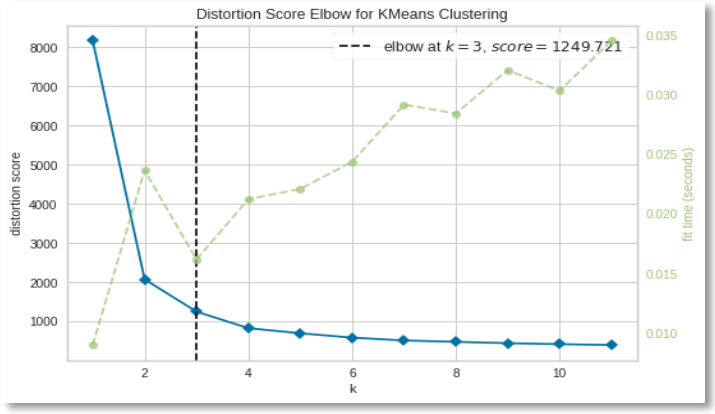
Kemudian diperoleh hasil perhitungan *SSE* dari *Python* dibawah ini :

```
[55] kmeans.inertia_

     1249.7209595959598
```

Gambar 23. Hasil perhitungan *SSE*

Maka didapatkan hasil *cluster* terbaik berjumlah K=3 yang membentuk siku seperti gambar dibawah ini :



Gambar 24. Hasil Metode *Elbow*

F. *Deployment*

Pada tahap ini *deployment* merupakan tahapan terakhir berupa pengetahuan atau informasi mengenai pola kehadiran pegawai Institut Teknologi Pagar Alam sehingga dapat diketahui ada 3 *cluster* yang pertama keterangan *cluster_0* adalah tepat waktu, kemudian *cluster_1* memiliki keterangan sedang, dan *cluster_2* dengan keterangan Tidak Tepat Waktu. Dari proses *clustering* ini pihak instansi terutama bagian kepegawaian bisa melihat perbandingan dari pola kehadiran pegawai selama 3 tahun.

4. KESIMPULAN

Berdasarkan penelitian yang telah dilakukan, diperoleh hasil bahwa penelitian ini menghasilkan Pola Klasterisasi Kehadiran Pegawai pada Institut Teknologi Pagar Alam. Hasil dari proses *clustering* data dalam mengetahui pola kehadiran pegawai selama tiga tahun melalui aplikasi *Rapid Miner* sama dengan hasil yang diterapkan di *python* dengan jumlah *cluster* yang terdiri dari 3 *cluster* yakni kehadiran pegawai ditahun 2020 yaitu *Cluster_0* memiliki tingkat kehadiran Tepat Waktu dengan Jumlah 6 Pegawai, *Cluster_1* memiliki tingkat kehadiran Sedang dengan jumlah 10 pegawai, *Cluster_2* memiliki tingkat kehadiran Tidak Tepat Waktu dengan jumlah pegawai 9. Selanjutnya ditahun 2021 diperoleh *cluster_0* sebanyak 6 pegawai dengan tingkat kehadiran Tepat Waktu, *cluster_1* sebanyak 11 pegawai dengan tingkat kehadiran sedang, dan *cluster_2* sebanyak 8 pegawai dengan keterangan Tidak Tepat Waktu. Dan ditahun 2022 diperoleh *cluster_0* sebanyak 15 pegawai dengan tingkat kehadiran Tepat Waktu, *cluster_1* sebanyak 8 pegawai dengan tingkat kehadiran sedang, dan *cluster_2* sebanyak 2 pegawai dengan keterangan Tidak Tepat Waktu. Dari hasil pola kehadiran ini dapat dikatakan dari 25 pegawai rata-rata kehadirannya adalah sedang yaitu dari jam 08.00-08.15 dan pulang <15.00 WIB, dari sini sudah didapatkan beberapa pengetahuan yang diharapkan bermanfaat bagi pihak instansi terutama bagian kepegawaian untuk melakukan tindakan selanjutnya. Kemudian nilai dari pengujian menggunakan *Elbow Method* yang berbentuk siku dengan menghitung hasil dari *Sum of Square Erorr (SSE)* diperoleh jumlah *cluster* yang tepat yaitu berjumlah K=3 dengan nilai 1249.721. maka dari hasil tersebut dapat dikatakan valid dengan hasil *clustering k-means* pada *rapid miner*.

ACKNOWLEDGEMENTS

Paper ini merupakan hasil dari penelitian tugas akhir mahasiswa Institut Teknologi Pagar Alam Tari Kristianda program studi Teknik Informatika, yang sedang proses skripsi.

DAFTAR PUSTAKA

[1] K. Handoko, “Penerapan Data Mining Dalam Meningkatkan Mutu Pembelajaran Pada Instansi Perguruan Tinggi Menggunakan Metode K - Means Clustering (Studi Kasus Di Program Studi Tkj Akademi Komunitas Solok Selatan),” vol. 02, no. 03, pp. 31–40, 2016.

[2] J. Informasi, W. Robiansyah, and G. W. Nurcahyo, “Ketepatan Pemberian Insentif Menggunakan Algoritma K-Medoids Pada Tingkat Kedisiplinan Pegawai,” vol. 3, 2021, doi: 10.37034/jidt.v3i3.125.

[3] N. F. Adani *et al.*, “Implementasi data mining untuk mengelompokkan data penjualan berdasarkan pola pembelian menggunakan algoritma clustering K-Means di toko Syihan,” no. x, pp. 1–11, 2019.

[4] I. Kusdinar, “Penerapan K-Means Clustering Untuk Penentuan Keputusan Pegawai Teladan Dan Berpotensiphk Bedasarkan Data,” Vol. 6, No. 1, Pp. 1–9, 2019.

[5] S. Informasi *Et Al.*, “Implementation Of Employee Discipline Clustering At Gotting Sidodadi Village Office Bandar Pasir Mandoge Using K-Means Kantor Desa Gotting Sidodadi Bandar Pasir Mandoge Menggunakan Algoritma K-Means,” vol. 3, no. 2, pp. 295–304, 2022.

[6] M. A. Hasanah, S. Soim, and A. S. Handayani, “Implementasi CRISP-DM Model Menggunakan Metode Decision Tree dengan Algoritma CART untuk Prediksi Curah Hujan Berpotensi Banjir,” vol. 5, no. 2, 2021.