PENERAPAN *DATA MINING* MENGGUNAKAN ALGORITMA *K-MEANS CLUSTERING* UNTUK MENENTUKAN LANGKAH DAN PROGRAM DALAM MENJARING SISWA BARU DI DALAM DAN DI LUAR LINGKUNGAN MI MIFTAHUL HUDA SUKOLILO-JABUNG

Oleh :

Nama : Dini Kristianti

NIM : 200605220015

Kelas : B

UNIVERSITAS ISLAM NEGERI

MAULANA MALIK IBRAHIM

MALANG

2021

Abstrak

Data Mining adalah proses menemukan hubungan yang berarti, pola, dan kecenderungan dengan memeriksa dalam sekumpulan besar data yang tersimpan dalam penyimpanan dengan menggunakan teknik pengenalan pola seperti teknik statistic dan matematika. Proses penjaringan dan penerimaan siswa baru di MI Miftahul Huda Sukolilo pada tiap tahunnya selalu menggunakan data berkas (kertas) yang dapat membuat hardfile menjadi menumpuk. Oleh karena itu untuk merapikan dan memudahkan dalam pencarian data, pada tugas ini akan dibuatkan penerapan dalam data mining, supaya hardfile yang ada dapat tersimpan dengan baik, rapi di dalam penyimpanan data yang besar.implementasi K-Means Clustering menggunakan RapidminerStudio 9.9. Atribut yang digunakan adalah asal kelurahan dan jumlah siswa yang berada dalam kelurahan tersebut.

Kata kunci: *Data Mining, K-Means, Clustering, Rapid Miner 9.9, Siswa*

DAFTAR ISI

PENDAHULUAN

Seiring berkembangnya kemajuan teknologi informasi yang banyak memberikan pengaruh besar pada dunia Pendidikan, baik pada tingkat Pendidikan Usia Dini, Taman Kanak –kanak, SD/MI, SMP/ MTS, SMA/SMK/MA dan pada tingkat perguruan tinggi. Hal ini terlihat dalam penggunaan komputer pada setiap pekerjaan di dunia Pendidikan. Penggunaan teknologi ini digunakan untuk mendukung kegiatan sehari hari yang di lakukan pada dunia pendidikan.

Mi Miftahul Huda Sukolilo setiap tahunnya memiliki data dalam bentuk hardfile dalam menerima siswa baru di sekolah. Data yang sangat banyak itu, kemudian di inputkan kedalam komputer, hasilnya data itu sangatlah besar dalam basis data.

Jumlah data yang terus meningkat memerlukan beberapa metode untuk mengolah dan mengidentifikasi data tersebut. Beberapa metode yang di gunakan untuk mengolah data yang sifatnya besar supaya dapat menemukan pola yang terdapat di dalamnya salah satunya adalah menggunakan metode *K-Means Clustering Data*.

Untuk dapat mengambil langkah dan membuat program dalam menjaring siswa baru supaya lebih baik dan efektif, serta menghemat biaya yang di keluarkan. Maka dalam tugas ini dilakukan dengan cara mengolah data-data untuk mengetahui pola dari data-data tersebut sehingga kita dapat mengambil informasi-informasi yang tersembunyi dari data-data tersebut. Metode pengolahan data seperti ini sering disebut sebagai data mining. Pada tugas ini analisa data mining dilakukan dengan menggunakan metode K-Means *Clustering*. Dengan menggunakan metode ini, data yang telah didapatkan dapat dikelompokan ke dalam beberapa cluster berdasarkan kemiripan dari data-data tersebut, sehingga data-data yang memiliki karakteristik yang sama dikelompokan dalam satu cluster dan yang memliki karakteristik yang berbeda dikelompokan dalam cluster yang lainyang memiliki karakteristik yang sama.Dengan adanya pengelompokan pengelompokan data seperti ini,diharapkan Panitia PPDB dapat menentukan langkah dan menyusun program yang tepat untuk mendapatkan calon siswa baru.

LANDASAN TEORI

Data Mining

Menurut Gartner Group, *data mining* adalah suatu proses menemukan hubungan yang berarti, pola, dan kecenderungan dengan memeriksa dalam sekumpulan besar data yang tersimpan dalam penyimpanan dengan menggunakan teknik pengenalan pola seperti teknik statistic dan matematika. Data mining bukanlah suatu bidang yang sama sekali baru. Salah satu kesulitan untuk mendefinisikan data mining adalah kenyataan bahwa *data mining* mewarisi banyak aspek dan teknik dari bidang-bidang ilmu yang sudah mapan terlebih dulu. Berawal dari beberapa disiplin ilmu, *data mining* bertujuan untuk memperbaiki teknik tradisional sehingga bias menangani:

1. Jumlah data yang sangat besar.
2. Dimensi data yang tinggi.
3. Data yang heterogen dan berbeda bersifat

Algoritma *K-Means Clustering*

Cluster Analysis merupakan salah satu metode *objek mining* yang bersifat tanpa latihan(*unsupervised analysis*), sedangkan *K-Means Cluster Analysis* merupakan salah satu metode *clusteranalysis* non hirarki yang berusaha untuk mempartisi objek yang ada kedalam satu atau lebih *cluster*atau kelompok objek berdasarkan karakteristiknya, sehingga objek yang mempunyai karakteristikyang sama dikelompokan dalam satu *cluster* yang sama dan objek yang mempunyai karakteristik yangberbeda dikelompokan kedalam *cluster* yang lain. Tujuan *pengelompokan* adalah untuk meminimalkan*objective function* yang di set dalam proses *clustering*, yang pada dasarnya berusaha untukmeminimalkan variasi dalam satu *cluster* dan memaksimalkan variasi antar *cluster*. Metode *cluster* ini meliputi *sequential threshold, pararel threshold dan optimizing threshold.Sequential threshold* melakukan pengelompokan dengan terlebih dahulu memilih satu objek dasar yang akan dijadikan nilai awal *cluster*, kemudian semua *cluster* yang ada dalam jarak terdekat dengan *cluster* ini akan bergabung, lalu dipilih *cluster* kedua dan semua objek yang mempunyai kemiripan dengan *cluster* ini akan digabungkan, demikian seterusnya sehingga terbentuk beberapa *cluster* dengan keseluruhan objek yang terdapat didalamnya.

Menurut (Santosa, 2007) dan Ong,langkah-langkah melakukan *Clustering*dengan metode K-Means adalah sebagaiberikut:

1. Pilih jumlah *cluster* k.

2. Inisialisasi k pusat *cluster* ini bisa dilakukan dengan berbagai cara.Namun yang paling sering dilakukan adalah dengan cara random. Pusat pusat *cluster* diberi nilai awal dengan angka-angka random,
3. Alokasikan semua data/ objek ke*cluster* terdekat. Kedekatan dua objek ditentukan berdasarkan jarak kedua objek tersebut. Demikian juga kedekatan suatu data ke cluster tertentu ditentukan jarak antara data dengan pusat cluster. Dalam tahap ini perlu dihitung jarak tiap data ke tiap pusat cluster. Jarak paling antara satu data dengan satu cluster tertentu akan menentukan suatu data masuk dalam *cluster* mana. Untuk menghitung jarak semua data ke setiap titik pusat cluster dapat menggunakan teori jarak *Euclidean* yang dirumuskan sebagaiberikut:

$$D(i,j) = \sqrt{(X1i - X1j)^2 + (X2i - X2j)^2 + \ldots + (Xki - Xkj)^2}$$

Dimana :

D(i,j) = Jarak data ke i ke pusat cluster j
Xki = Data ke i pada atribut data ke k
Xkj = Titik pusat ke j pada atribut ke k

4. Hitung kembali pusat *cluster* dengan keanggotaan *cluster* yang sekarang.Pusat *cluster* adalah rata-rata dari semua data/objek dalam *cluster* tertentu. Jika dikehendaki bisa juga menggunakan median dari *cluster* tersebut. Jadi rata-rata (mean) bukan satu-satunya ukuran yang bisa dipakai
5. Tugaskan lagi setiap objek memakai pusat *cluster* yang baru. Jika pusat *cluster* tidak berubah lagi maka proses *Clustering* selesai. Atau, kembali kelangkah nomor 3 sampai pusat *cluster* tidak berubah lagi.

RapidMiner

Rapid miner adalah aplikasi *data mining* yang berbasis *open source. Open source rapid miner* berlisensi AGPL (*GNU Affero General Public License)* versi 3. Penelitian mengenai *tools* ini dimulai sejak tahun 2001 oleh Ralf Klinkenberg, Ingo Mierswa, dan Simon Fischer di *Artificial Intelligence Unit* dari University of Dortmund yang kemudian diambil alih oleh SourceForge sejak tahun 2004. Rapid miner memperoleh peringkat satu sebagai *tools data mining* untuk proyek nyata pada poll oleh KDnuggets, sebuah koran *datamining* pada 2010-2011. Dalam penerapannya, rapid miner menyediakan prosedur *data mining* dan *machine learning* termasuk : ETL (*extraction, transformation, loading*), data *preprocessing,* visualisasi, *modelling* dan evaluasi. Proses *data mining* tersusun atas operator-operator yang

nestable, dideskripsikan dengan XML, dan dibuat dengan GUI. *Tools* rapid miner ditulis dalam bahasar pemrograman Java dan juga mengintegrasikan proyek *data mining* Weka dan statistika. Beberapa solusi yang diusung oleh rapid miner antara lain :

a. Integrasi data, Analitis ETL, Data Analisis, dan pelaporan dalam satu suite tunggal.
b. Powerfull tetapi memiliki antarmuka pengguna grafis yang intuitif untuk desain anakisis proses.
c. Repositori untuk prose, data dan penanganan meta data.
d. Hanya solusi dengan transformasi meta data: lupakan trail and arror dan memeriksa hasil yang telah diinspeksi selama desain.
e. Hanya solusi yang mendukung *on-the-fly* kesalahan dan dapat melakukan perbaikan dengan cepat. Lengkap dan fleksibel: ratus an *loading* data, transformasi data, pemodelan data dan metode visualisasi data.

METODE PENELITIAN

Metode yang digunakan adalah metode kualitatif dengan pendekatan deskriptif. Metode kualitatif sering disebut metode penelitian naturalistik karena penelitiannya dilakukan pada kondisi yang alamiah, dapat diartikan sebagai usaha untuk menyelidiki keadaan yang sebenarnya, dalam memprediksi prilaku pola pembelian berdasarkan jenis produk sehingga perencanaan strategi penjualan dapat tercapai.

HASIL DAN PEMBAHASAN

Proses Data Mining

Persiapan Data

Dari 420 data siswa pada Tahun Ajaran 2020/ 2021, maka dilakukan teknik data preparation agar kualitas data diperoleh lebih baik dengan cara :

1. Data validation, menghapus data dan mengidentifikasi data yang tidak konsisten dan data yang tidak lengkap
2. Data Itegration dan Transformatuon, meningkatkan akurasi dan efisisensi algoritma. Pada Kelurahan asal (Tempat Tinggal) dilakukan perhitungan data.

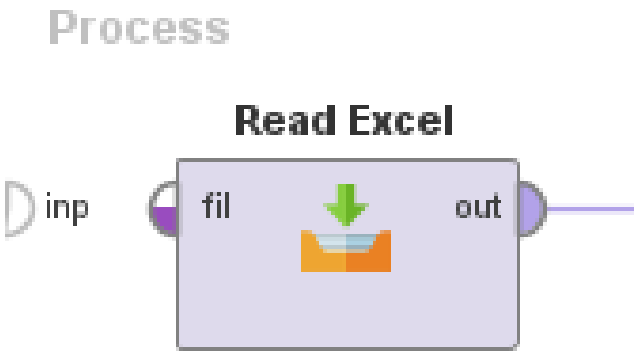
Berikut adalah data training dalam bentuk Excell yang akan di olah.

KELAS	JABUNG	SUKOLILO	KEMANTREN	KEMIRI	SIDOMULYO	ARGOSARI	PAKISJAJAR	LOWOKWARU	GADINGKEMBAR	BRANGKAL	TOTAL
1	5	19	17	2	1	1	10	1	2	1	59
2	12	40	12	1	0	0	0	0	3	0	68
3	2	36	27	0	4	3	5	0	5	0	82
4	3	20	38	1	1	1	3	0	3	0	70
5	1	20	14	1	1	6	4	0	3	0	50
6	16	30	23	5	2	4	5	0	6	0	91

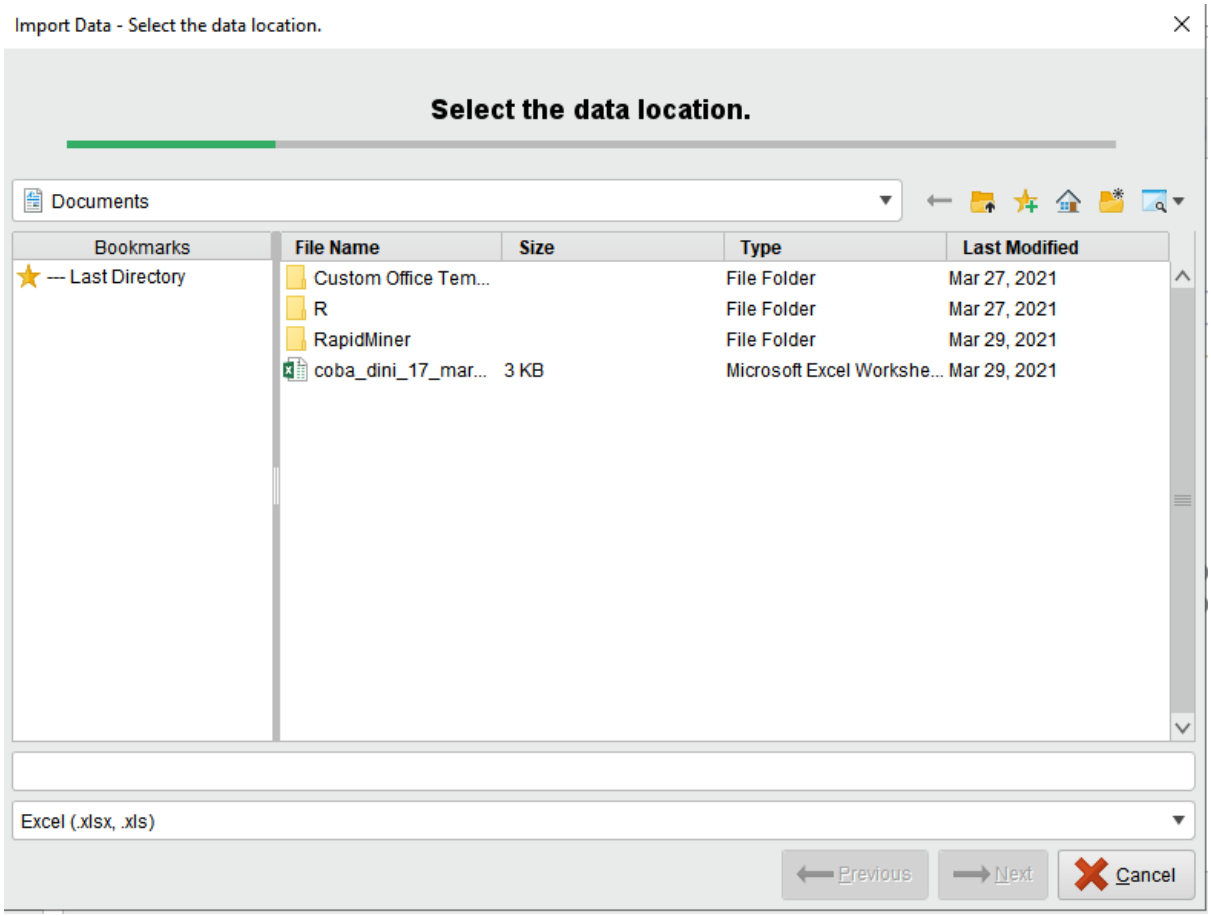
Gambar 1 Data Jumlah Siswa Tiap Kelas Pada Setiap Kelurahan dalam format Excell

Proses Training dan Testing

Pada proses Training dan Testing, silahkan di mulai dengan operator Read Excell, setelah operator Read Excell sudah di drag pada lemabr kerja, silahkann klik 2x (Gambar 2), dan cari direktori data yang akan di proses (gambar 3)

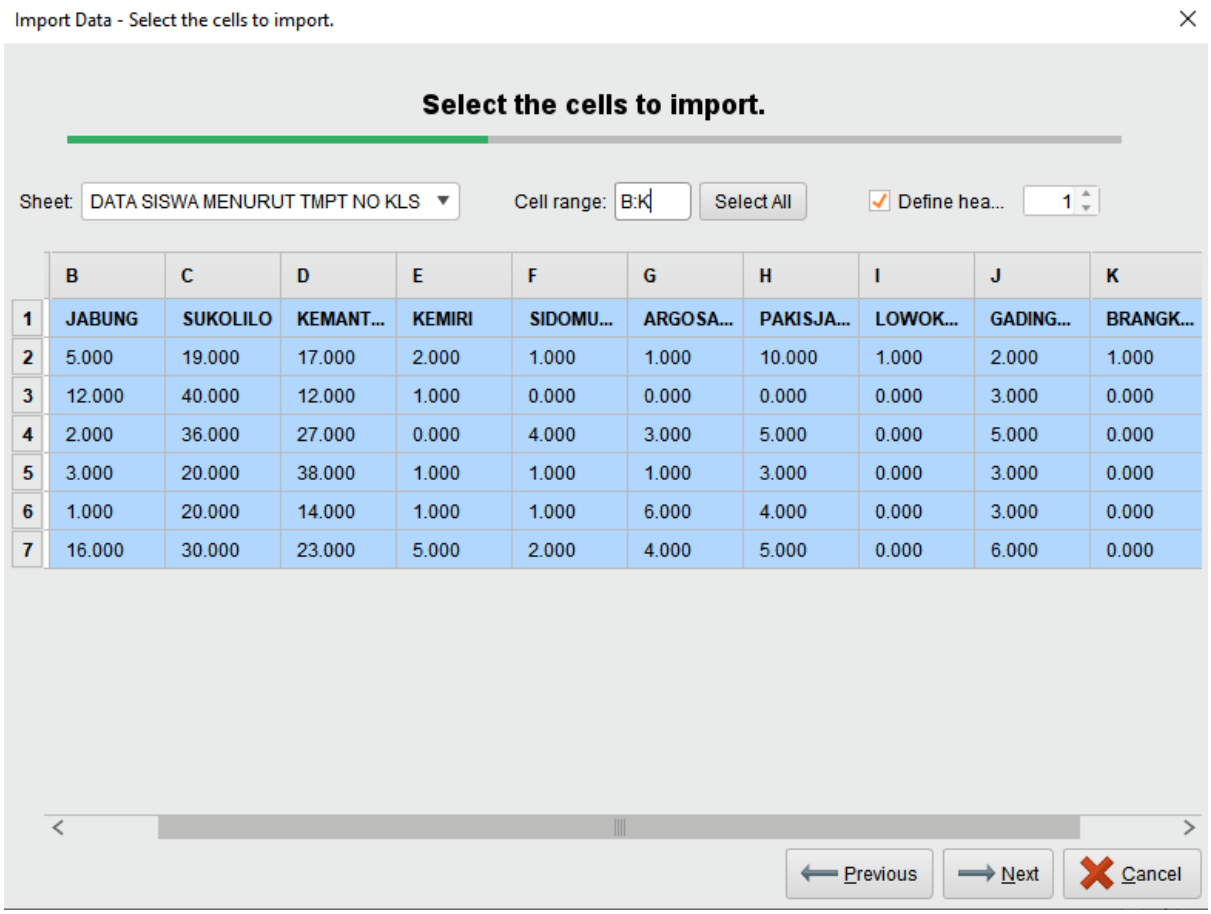


Gambar 2



Gambar 3

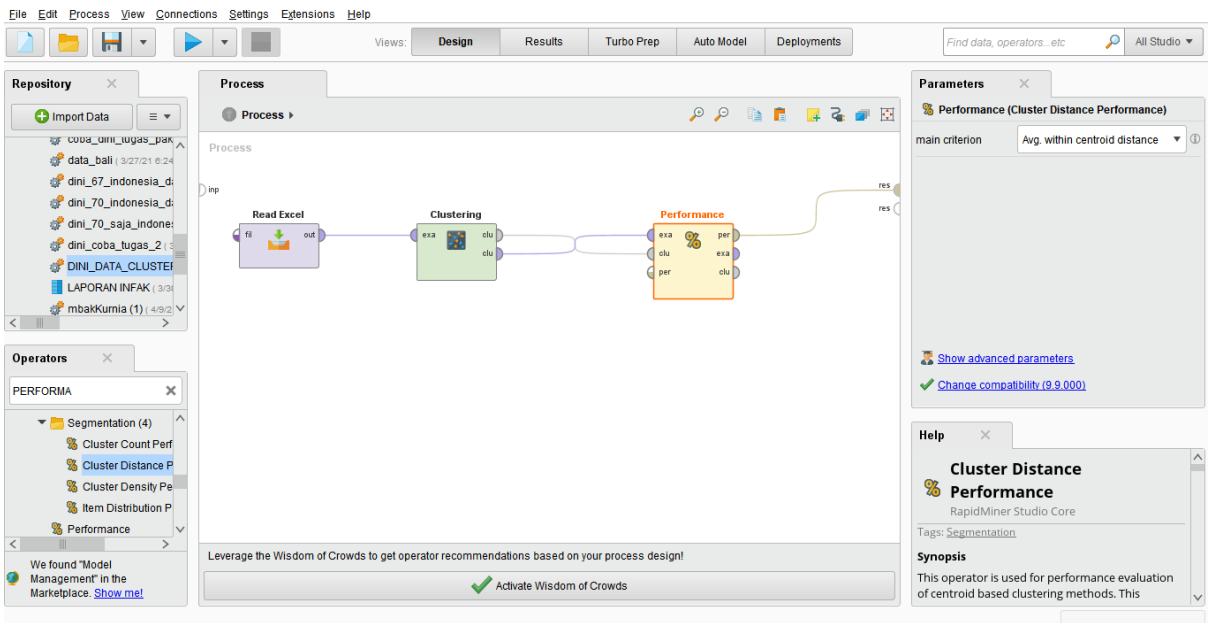
Setelah di temukan data yang akan di proses, langkah selanjutnya yaitu isi range cell yang akan di buat clustering, lihat pada gambar 4.



	B	C	D	E	F	G	H	I	J	K
1	JABUNG	SUKOLILO	KEMANT...	KEMIRI	SIDOMU...	ARGOSA...	PAKISJA...	LOWOK...	GADING...	BRANGK...
2	5.000	19.000	17.000	2.000	1.000	1.000	10.000	1.000	2.000	1.000
3	12.000	40.000	12.000	1.000	0.000	0.000	0.000	0.000	3.000	0.000
4	2.000	36.000	27.000	0.000	4.000	3.000	5.000	0.000	5.000	0.000
5	3.000	20.000	38.000	1.000	1.000	1.000	3.000	0.000	3.000	0.000
6	1.000	20.000	14.000	1.000	1.000	6.000	4.000	0.000	3.000	0.000
7	16.000	30.000	23.000	5.000	2.000	4.000	5.000	0.000	6.000	0.000

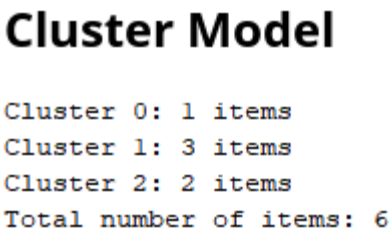
Gambar 4

Jika data sudah berhasil di import, silahkan berikan operator *K-Means* supaya kita bisa mengetahui hasil dari clustering data kita. Disini clusternya, di buat 3.



Gambar 5

Dengan menggunakan pemodelan *K-Means Clustering* seperti Gambar 5 diatas, dengan inisialisasi jumlah cluster sebanyak 3, maka didapat hasil sebagai berikut :



Gambar 6 Hasil Clustering

Cluster 0 adalah jumlah Kelurahan yang paling sedikit siswanya, cluster 1 adalah jumlah kelurahan ke 2 jumlah siswanya

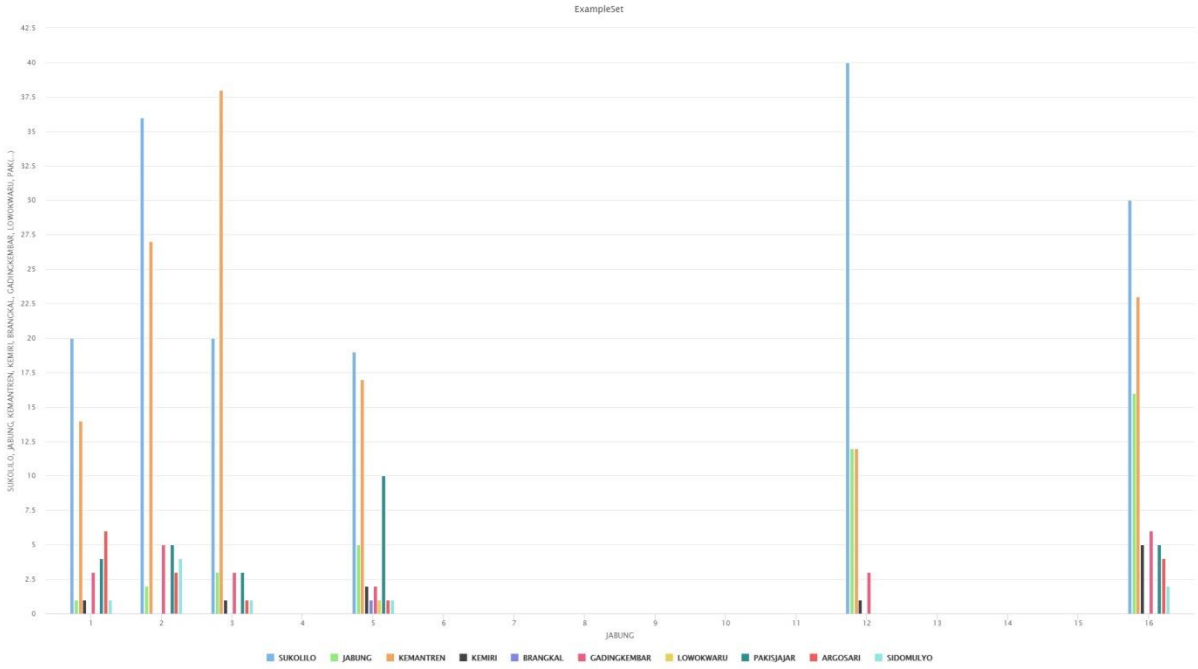
cluster	JABUNG	SUKOLILO	KEMANTREN	KEMIRI	SIDOMULYO	ARGOSARI	PAKISJAJAR	LOWOKWA...	GADINGKEM...	BRANGKAL
cluster_2	5	19	17	2	1	1	10	1	2	1
cluster_1	12	40	12	1	0	0	0	0	3	0
cluster_1	2	36	27	0	4	3	5	0	5	0
cluster_0	3	20	38	1	1	1	3	0	3	0
cluster_2	1	20	14	1	1	6	4	0	3	0
cluster_1	16	30	23	5	2	4	5	0	6	0

Gambar 7

Attribute	cluster_0	cluster_1	cluster_2
JABUNG	3	10	3
SUKOLILO	20	35.333	19.500
KEMANTREN	38	20.667	15.500
KEMIRI	1	2	1.500
SIDOMULYO	1	2	1
ARGOSARI	1	2.333	3.500
PAKISJAJAR	3	3.333	7
LOWOKWARU	0	0	0.500
GADINGKEMBAR	3	4.667	2.500
BRANGKAL	0	0	0.500

Gambar 8

Hasil Analisis Cluster menggunakan Chart:



Gambar 9

Berdasarkan Chart di atas dapat dilihat 2 Kelurahan yang banyak jumlah siswanya dari 6 kelas yaitu Sukolilo pada tingkat pertama dan pada tingkat kedua yaitu Kemantren.

SIMPULAN DAN SARAN

Setelah dilakukan pengelompokan data siswa melalui persebran Kelurahan (Tempat tinggal) menggunakan *K-Means Clustering*, maka dapat disimpulkan sebagai berikut :

1. Setelah dilakukan pengelompokan data siswa melalui persebran kelurahan tempat tinggal siswa menggunakan K-Means clustering terbentuk tiga cluster yaitu cluster 2 dengan jumlah 2 items, cluster 1 sebanyak 3 item dan cluster 0 sebanyak 1 items.
2. Dari hasil clustering data di atas, maka yang di perlu dilakukan oleh panitia PPDB adalah menyiapakan tim Panitia PPDB yang sangat professional untuk mendekatkan diri ke Sekolah TK yang berada di wilayah kelurahan yang banyak terdapat siswanya.

DAFTAR PUSTAKA

Ediyanto, Mara, M. N., & Satyahade, N. (2013, 1-4). PENGKLASIFIKASIAN KARAKTERISTIK DENGAN METODE K-MEANS CLUSTER ANALYSIS. *Buletin Ilmiah Mat. Stat. dan Terapannya (Bimaster), 2*.

SABNA, E. (2017). ANALISIS DATA MAHASISWA DENGAN ALGORITMA K-MEANS UNTUK MENDUKUNG STRATEGI PROMOSI STIKES HANG TUAH PEKANBARU. *JURNAL ILMU KOMPUTER, 6*, 1-6.

Santosa, B. (2007). Data mining teknik pemanfaatan data untuk keperluan bisnis. *978*(979), 756.

Setiawan , R. (2016). PENERAPAN DATA MINING MENGGUNAKAN ALGORITMA K-MEANS CLUSTERING UNTUK MENENTUKAN STRATEGI PROMOSI MAHASISWA BARU (Studi Kasus : Politeknik LP3I Jakarta). *JURNAL LENTERA ICT, 3*, 1-17.