Faktor Exacta

Vol. 15, No. 2, June 2022, pp. 127~138
eISSN: 2502-339X, pISSN: 1979-276X, DOI: https://doi.org/10.30998/faktorexacta.v15i2.13211

MODEL *MACHINE LEARNING* KLASIFIKASI DATA SEKOLAH TK BERDASARKAN STATUS DAN KABUPATEN/KOTA ADMINISTRASI PROVINSI DKI JAKARTA

Abdurahman[1], Fiqih Ismawan[2]
[1]Department of Informatic, Universitas Indraprasta PGRI, Indonesia
[2]Department of Informatic, Universitas Indraprasta PGRI, Indonesia

Article Info

Article history:

Received June 10, 2022
Revised July 14, 2022
Accepted July 29, 2022

Keywords:

Klasifikasi
Status Sekolah
K-Means Clustering
Levenshtein Distance
Machine Learning

ABSTRAK

Klasifikasi status sekolah menjadi parameter khusus bagi beberapa kalangan orang tua dalam melakukan pemilihan sekolah untuk anak-anaknya yang dinginkan, beberapa permasalahan orang tua dalam mempertimbangkan penentuan sekolah salah satunya adalah status sekolah, jumlah sekolah, jumlah guru, jumlah murid dan jumlah ruang kelas. Makalah ini melaporkan bahwa data status sekolah TK kabupaten dan kota administrasi provinsi DKI Jakarta dapat dilakukan klasifikasi berdasarkan *cluster* dan domain data, dengan mempartisi data ke dalam *cluster*. Karakteristik data yang sama dapat dikelompokkan ke dalam satu *cluster* yang sama dan data yang mempunyai karateristik yang berbeda dikelompokan ke dalam *cluster* yang lain. Sehingga dapat memudahkan masalah orang tua dalam menentukan sekolah yang diinginkan sesuai dengan parameter data yang dimiliki. Metode klasifikasi yang digunakan adalah *Levenshtein Distance dan K-Means Clustering*, sumber data yang digunakan dalam penelitian ini adalah data sekunder yang diperoleh data.jakarta.go.id. Data sekunder yang digunakan adalah data sekolah dari 12 *record* kabupaten dan kota di Jakarta. Penelitian ini bertujuan untuk memodelkan data dan menentukan kriteria sekolah serta menganalisis akurasi klasifikasi sekolah yang sesuai dengan parameter yang orang tua inginkan dengan menggunakan kedua metode tersebut dalam klasifikasi pemilihan data sekolah TK berdasarkan status dan kabupaten/kota administrasi provinsi DKI Jakarta. Setelah dilakukan pengujian maka hasil *Silhouette Score* berdasarkan *Average* dari 4 atribut yaitu Cluster C1 dari *score* 0,691355 sampai 0,718406, Cluster C2 dari *score* 0,745171 sampai 0,747778 dan Cluster C3 dari *score* 0,601115 sampai 0,647377. Hasil Penelitian ini berupa pemodelan data dengan menggunakan parameter yang diambil dari data.jakarta.go.id kemudian diuji menggunakan beberapa model klasifikasi yang terdapat pada *Machine Learning*.

Corresponding Author:

Fiqih Ismawan, Department of Informatic,
Universitas Indraprasta PGRI,
Jl. Raya Tengah No. 80, Jakarta Timur 13760, Daerah Khusus Ibukota Jakarta 12530
Email: vQ.unindra@gmail.com

1. PENDAHULUAN

Sesuai dengan visi misi dinas pendidikan DKI Jakarta bahwa DKI Jakarta ingin mewujudkan pendidikan yang tuntas dan berkualitas untuk semua, maka peran sekolah menjadi parameter utama dalam mewujudkan pendidikan yang berkualitas. Pendidikan Tingkat Kanak-Kanak (TK) dan sederajatnya merupakan jenjang paling dasar pada pendidikan formal yang mempunyai peran besar bagi keberlangsungan proses pendidikan selanjutnya untuk anak. Bagi siswa tingkat kanak-kanak yang ingin melanjutkan pendidikannya kejenjang yang lebih tinggi, maka anak harus mengikuti aturan yang telah ditetapkan oleh pemerintah. Untuk PPDB Jakarta, ada persyaratan usia yang harus dipenuhi Calon Peserta Didik Baru (CPBD), Persyaratan usia tersebut diatur dalam Peraturan Gubernur (Pergub) Nomor 32 Tahun 2021 tentang Petunjuk Teknis PPDB, tepatnya di Pasal 4 Pergub Nomor 32 Tahun 2021. Adapun usia minimal masuk PAUD PPDB Online Tahun Ajaran 2021/2022 adalah usia minimal

Taman Penitipan Anak dan Satuan PAUD sejenis: 2-6 tahun pada tanggal 1 Juli tahun berjalan, usia minimal Kelompok Bermain (KB): 3-4 tahun pada tanggal 1 Juli tahun berjalan, usia minimal kelompok A Taman Kanak-Kanak (TK): paling rendah berusia 4 tahun pada tanggal 1 Juli tahun berjalan, usia minimal kelompok B Taman Kanak-Kanak (TK): paling rendah berusia 5 tahun pada tanggal 1 Juli tahun berjalan. Dari hasil uji kelayakan tersebut akan menentukan masuk dan tidaknya siswa dari TK ke Sekolah Dasar (SD). Dalam setiap tahunnya setiap sekolah akan menyetorkan laporan individu tentang sekolahnya pada Dinas Pendidikan, yang juga terdapat nilai hasil rata-rata dari seluruh siswa yang masuk uji kelayakan tersebut.

Klasifikasi status sekolah menjadi parameter khusus bagi beberapa kalangan orang tua dalam melakukan pemilihan sekolah untuk anak-anaknya yang dinginkan, beberapa permasalahan orang tua dalam mempertimbangkan penentuan sekolah salah satunya adalah status sekolah, jumlah sekolah, jumlah guru, jumlah murid dan jumlah ruang kelas. Penelitian ini bermaksud untuk melaporkan bahwa data status sekolah TK kabupaten dan kota administrasi provinsi DKI Jakarta dapat dilakukan olah data pada klasifikasi berdasarkan *cluster* dan domain data tiap-tiap sekolah TK yang ada di kabupaten dan kota administrasi provinsi DKI Jakarta, dengan cara mempartisi data ke dalam beberapa *cluster*. Kemudian karakteristik data yang sama nantinya dapat dikelompokkan ke dalam satu *cluster* yang sama dan data yang mempunyai karateristik yang berbeda nantinya dapat dikelompokan ke dalam cluster yang lain. Sehingga dapat memudahkan permasalahan orang tua dalam menentukan sekolah yang diinginkan sesuai dengan parameter data yang dimiliki. Metode klasifikasi yang digunakan dengan menerapkan beberapa model *Machine Learning* (ML), salah satu teknik data mining yang bersifat *unsupervised learning* seperti *Levenshtein Distance* dan *K-Means Clustering.* Sumber data yang digunakan sebagai acuan adalah data sekunder yang diperoleh data.jakarta.go.id, data sekunder yang digunakan adalah data sekolah dari 12 *record* kabupaten dan kota di Jakarta. Penelitian ini bertujuan untuk memodelkan data dan menentukan kriteria sekolah serta menganalisis akurasi klasifikasi sekolah yang sesuai dengan parameter yang orang tua inginkan dengan menggunakan kedua metode tersebut dalam klasifikasi pemilihan data sekolah TK berdasarkan status dan kabupaten/kota administrasi provinsi DKI Jakarta

Machine Learning (ML) atau pembelajaran mesin merupakan pendekatan dalam *Artificial Intelligence* (AI) [1]. ML menjadi salah satu bidang ilmu komputer yang tumbuh paling cepat, dengan aplikasi yang luas jangkauannya [2][3]. Algoritma *machine learning* memiliki beberapa jenis diantaranya *supervised learning algorithms*, *unsupervised learning algorithms*, *semi-supervised learning* dan *reinforcement learning* [4][5][6]. Penelitian ini berfokus pada salah satu algoritma *machine learning* yaitu *unsupervised learning*. *Unsupervised learning* adalah algoritma *machine learning* yang digunakan untuk menarik simpulan dari dataset [7]. Metode ini hanya akan mempelajari suatu data berdasarkan jalur yang terdekat atau yang biasa disebut dengan *clustering* [8]. Cara kerja *unsupervised learning* yang paling umum adalah menganalisis *cluster* data untuk mencari pola-pola tersembunyi atau pengelompokan dalam data [9].

Algoritma *unsupervised learning* yang menjadi fokus pada penelitian ini adalah *Levenshtein Distance (LD)* dan *K-Means Clustering.* LD merupakan algoritma yang mengukur kesamaan antara 2 string yakni string sumber (s) dengan string target (t), algoritma ini ditemukan oleh Vladimir Losifovich Levenshtein yang merupakan ilmuan dari Rusia pada tahun 1965 [10]. Cara kerja algoritma ini yaitu mensimulasikan data melalui model matriks. *K-Means Clustering* adalah salah satu cara pengelompokkan data sekatan (nonhierarki) yang berusaha membagi data yang ada kedalam bentuk dua atau lebih kelompok [11][12]. Teknik pengelompokan ini bekerja berdasarkan partisi *cluster*, dan pada setiap iterasi dari hasil pengelompokan hirarki *cluster* [13][14].

Beberapa penelitian yang menggunakan metode *clustering* dalam melakukan klasifikasi data diantaranya, penelitian yang dilakukan oleh Gustientiedina dkk (2019), penelitian ini menerapkan Algoritma K-Means Untuk *Clustering* data obat-obatan pada RSUD Pekanbaru dengan hasil klasterisasi pada data obat – obatan ini bahwa kelompok pemakaian sedikit, sedang dan tinggi [15]. Penelitian yang dilakukan oleh Novianti dkk (2017) menerapkan algoritma *K-Means Clustering* dengan metode *Euclidean distance* dalam analisis data *clustering* gempa tektonik di Provinsi Bengkulu dan sekitarnya dari Januari 1970 sampai Desember 2015 dengan variabelnya adalah garis lintang, garis bujur dan besaran [14]. Penelitian yang dilakukan Michael Chau dkk (2006) menerapkan algoritma *K-means Clustering* pada pola tertentu dari ketidakpastian suatu objek bergerak. Hasil eksperimen menunjukkan bahwa dengan mempertimbangkan ketidakpastian, algoritma *clustering* dapat menghasilkan hasil yang lebih akurat [12]. Penelitian yang dilakukan oleh Indriyani dkk (2019) dengan membuat *clustering* pengelompokan data penjualan agar dapat memaksimalkan manajemen stok barang pada toko retail di kota Bogor yang menjual peralatan *outdoor* Genta Corp. Hasil penelitian berupa tiga *cluster* yaitu terdapat 2 jenis barang paling laris, 8 jenis barang yang cukup laris dan 18 jenis barang yang kurang laris [11]. Penelitian yang dilakukan oleh Kiki Fatmawati dkk (2018) melakukan pengelompokkan jumlah daerah yang terjangkit demam berdarah dengue (DBD) berdasarkan provinsi, metode yang digunakan adalah Data mining *K-Means Clustering.* Penelitian ini menggunakan sumber data yang terekam di situs kementerian kesehatan https://www.depkes.go.id/, data yang digunakan adalah (Tahun 2014-2016) yang terdiri dari 34 provinsi [13]. Samit Ghosal dkk (2020) melakukan analisis regresi sederhana untuk menilai tingkat perubahan infeksi dan tingkat kematian pada penyebaran infeksi SARS-CoV-2, dengan menganalisis menggunakan algoritma k-means dan klaster hierarki untuk mengidentifikasi negara-negara yang memiliki kinerja serupa dalam pemberlakuan *lockdown* guna

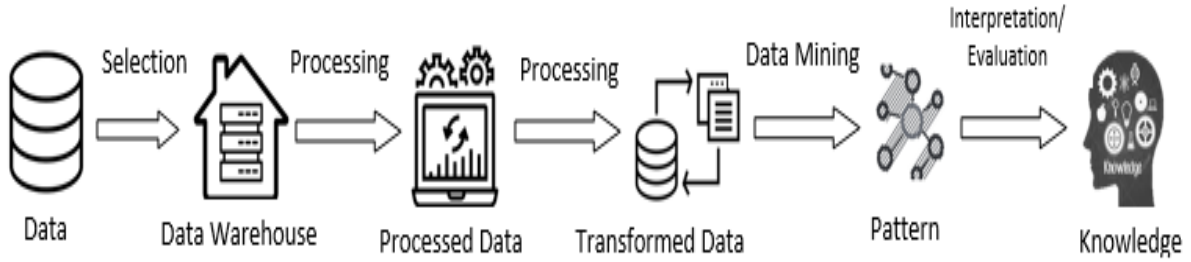
penanganan penyebaran infeksi SARS-CoV-2 yang terdapat di negara-negara tertentu [16]. Gueddah Hicham dkk (2012) menyajikan pendekatan baru yang didedikasikan untuk mengoreksi kesalahan ejaan bahasa arab, pendekatan yang dilakukan mengoreksi kesalahan ketik ejaan seperti memasukkan, menghapus, dan permutasi dengan algoritma *Levenshtein distance* [10].

2. METODE

Objek penelitian ini menggunakan data sekunder yang diperoleh data.jakarta.go.id, data sekunder yang digunakan adalah data sekolah dari 12 *record* kabupaten dan kota di Jakarta.

2.1 Knowledge Discovery in Database (KDD)

Knowledge Discovery in Database (KDD) adalah suatu ilmu dari percabangan ilmu statistik, *database*, AI, visualisasi dan komputer paralel yang mempengaruhi pengetahuan *"interdisciplinary knowledge"* [5] yang melibatkan hasil ekstraksi kecenderungan suatu pola, sehingga mengubah hasilnya secara tepat dan akurat serta menjadi informasi yang mudah dipahami [17]. Langkah-langkah *Knowledge Discovery in Database* (KDD) pada penelitian ini terdapat pada gambar 1.



Gambar 1. Langkah *Knowledge Discovery in Database* (KDD)

Data *Selection*

Tahapan ini menetapkan variabel yang akan digunakan dalam proses data mining, dataset yang digunakan dalam penelitian ini adalah data yang tersedia di data.jakarta.go.id diakses dan dilakukan pemantauan data pada tanggal 20 Maret 2022 Pukul 10.00 wib, data tersebut berisi kumpulan data jumlah sekolah, guru, dan murid TK menurut status sekolah di DKI Jakarta, data yang dikumpulkan data pada tahun 2011 sampai sekarang untuk dimodelkan pada klasifikasi data menggunakan model *machine learning*.

Preprocessing

Tahapan ini perlu dilakukan proses *cleaning* dengan tujuan untuk membuang duplikasi data, memeriksa data yang inkonsisten dan memperbaiki kesalahan pada data seperti kesalahan cetak (tipografi). Juga dilakukan proses *enrichment*, yaitu proses "memperkaya" data yang sudah ada dengan data atau informasi lain yang relevan dan diperlukan untuk KDD, seperti data atau informasi eksternal.

Transformation

Tahapan ini memproses pengkodean pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses *data mining*. Proses *coding* dalam KDD merupakan proses kreatif dan sangat tergantung pada jenis atau pola informasi yang akan dicari dalam *database.*

Data *Mining*

Tahapan ini memproses serta mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Penelitian ini menerapkan teknik *clustering* yaitu metode *Levenshtein Distance* dan *K-Means Clustering*.

Pada penelitian ini, algoritma *Levenshtein Distance* pada dasarnya akan menghitung jumlah minimum dari upaya transformasi suatu string menjadi string lain. Transformasi ini meliputi penggantian, penghapusan, dan penyisipan. Berikut adalah operasi-operasi yang terdapat pada algoritma *Levenshtein Distance* seperti, operasi penyisipan karakter *(Insertion)*, operasi penghapusan karakter *(Deletion)* dan operasi penukaran karakter *(Subtitution)*. Terdapat beberapa langkah dalam menerapkan metode algoritma *Levenshtein Distance* [10]:

Langkah pertama (Penghapusan)

$$D(s,t) = \min D(s-1, t) + 1 \quad (1)$$

Langkah Kedua (Penyisipan)

$$D(s,t) = \min D(s, t-1) + 1 \quad (2)$$

Langkah Ketiga (Penggantian)

$$D(s,t) = \min D(s-1, t-1) + 1, sj \neq ti \quad (3)$$

Langkah Keempat (Tidak Ada Perubahan)

$$D(s,t) = \min D(s-1, t-1), sj = ti \quad (4)$$

Keterangan:
s: string sumber
D: jarak edit *Levenshtein Distance*
t(i): karakter string sumber ke-i
t: string target
s(j): karakter string sumber ke-j

Langkah awal pada algoritma *Levenshtein Distance*, yaitu melakukan penyeleksian panjang kedua string terlebih dahulu. Jika salah satu atau kedua string tersebut string kosong, maka jalannya algoritma ini berhenti dan memberikan hasil *edit distance* bernilai nol atau panjang string yang tidak kosong. Jika panjang string keduanya tidak nol, setiap string memiliki sebuah karakter terakhir, misalnya c1 dan c2. Misalnya bagian string pertama tanpa c1 adalah s1 dan bagian string kedua tanpa c2 adalah s2, dapat dikatakan penghitungan yang dilakukan adalah cara mentransformasikan s1+c1 menjadi s2+c2.

Jika c1 sama dengan c2, dapat diberikan nilai cost 0 dan nilai edit distance-nya adalah nilai edit *distance* dari pentransformasian s1 menjadi s2. Jika c1 berbeda dengan c2, dibutuhkan pengubahan c1 menjadi c2 sehingga nilai cost-nya 1. Akibanya, nilai edit distance-nya adalah nilai edit *distance* dari pentransformasian s1 menjadi s2 ditambah 1. Kemungkinan lain adalah dengan menghapus c1 dan mengedit s1 menjadi s2+c2 sehingga nilai edit *distance* dari pentransformasian s1 menjadi s2+c2 ditambah 1. Begitu pula dengan penghapusan c2 dan mengedit s1+c1 menjadi s2. Dari kemungkinan-kemungkinan tersebut, dicarilah nilai minimal sebagai nilai edit *distance*.

Kemudian pada penelitian ini menggunakan metode *K-Means Clustering*, Pada fase ini dilakukan pemilihan model yang akan digunakan untuk melakukan pengelompokan kumpulan data jumlah sekolah, guru, dan murid TK di DKI Jakarta berdasarkan status sekolah. Model atau metode yang akan digunakan pada penelitian ini adalah metode *K-Means Clustering*. Penerapan *K-Means Clustering* ini dapat dilakukan dengan prosedur *step by step* berikut [18]:
1. Siapkan data training berbentuk *vector*.
2. Set nilai K cluster.
3. Set nilai awal *centroids*.
4. Hitung jarak antara data dan centroid menggunakan rumus *Euclidean Distance.*
5. Partisi data berdasarkan nilai *minimum.*
6. Kemudian lakukan iterasi selama partisi data masih bergerak (tidak ada lagi objek yang bergerak ke partisi lain), bila masih maka ke poin 3.
7. Bila grup data sekarang sama dengan grup data sebelumnya, maka hentikan iterasi.
8. Data telah dipartisi sesuai nilai centroid akhir.

Berikut ini adalah rumus dari *Euclidean Distance*:

$$(x, y), (a, b) = \sqrt{(x-a)^2 + (y-b)^2} \quad (5)$$

Keterangan:
x = atribut x
y = atribut y
a = titik *centroid* a
b = titik *centroid* b.

Interpretation / Evaluation

Pola informasi yang dihasilkan dari proses data mining perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan. Tahap ini merupakan bagian dari proses KDD yang disebut dengan *interpretation.* Tahap ini mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesis yang ada sebelumnya atau tidak.

Pengujian akan dilakukan dengan membandingkan pengelompokan data yang dilakukan oleh *Levenshtein Distance* dan *K-Means Clustering*. Mengidentifikasi pola – pola yang menarik dalam *knowledge base* yang diidentifikasi. Pada tahap ini, menghasilkan pola – pola khusus maupun model prediksi yang dievaluasi untuk menilai kajian yang ada sudah memenuhi target yang diinginkan.

Knowledge

Pola yang dihasilkan dalam bentuk representasi kepada *user* sebagai pengetahuan baru untuk pengambilan keputusan.

3. HASIL DAN PEMBAHASAN

3.1 Hasil Data *Selection*

Data-data dikumpulkan dari data.jakarta.go.id yang diakses dan dilakukan pemantauan data pada tanggal 20 Maret 2022 Pukul 10.00 wib dapat dilihat pada tabel 1, berikut data yang terpantau secara *real time* berbentuk format csv. Data tersebut kumpulan data jumlah sekolah, guru, dan murid TK menurut status sekolah di DKI Jakarta, memiliki 6 atribut dan 12 *record* untuk prediksi dan klasifikasi menggunakan pemodelan algoritma *Levenshtein Distance* dan *K-Means Clustering*.

Tabel 1. Dataset Hasil *Data Selection* Untuk Prediksi dan Klasifikasi

Kabupaten Kota	Jenis Sekolah	Jumlah Sekolah	Jumlah Guru	Jumlah Murid	Jumlah Ruang Kelas
Kepulauan Seribu	Negeri	0	0	0	0
Kepulauan Seribu	Swasta	10	44	119	12
Jakarta Selatan	Negeri	7	43	371	20
Jakarta Selatan	Swasta	437	2500	19186	1448
Jakarta Timur	Negeri	4	24	187	8
Jakarta Timur	Swasta	0	2929	23368	1702
Jakarta Pusat	Negeri	608	30	238	11
Jakarta Pusat	Swasta	5	929	7135	523
Jakarta Barat	Negeri	208	0	0	0
Jakarta Barat	Swasta	435	2479	21138	1403
Jakarta Utara	Negeri	1	6	40	3
Jakarta Utara	Swasta	259	1464	11849	779

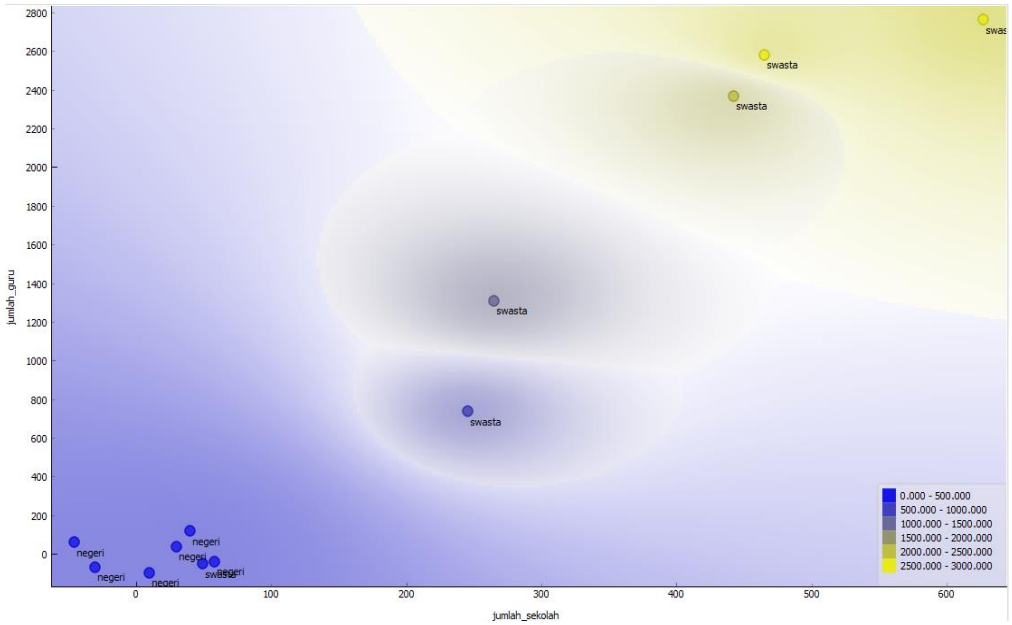
3.2 Hasil Preprocessing Data

Setelah dilakukan *data selection*, maka dataset pada tabel 1 dilakukan proses *preprocessing* data dengan teknik *impute* yaitu menghubungkan setiap data berdasarkan *average* atau *most frequency* pada setiap atribut jumlah sekolah, guru, murid dan jumlah ruang kelas secara *default* menggunakan value 0,000 seperti pada tabel 2. *Average* merupakan jarak rata-rata antara di dalam suatu *cluster* dengan elemen lainnya di dalam *cluster* yang berbeda.

Tabel 2. Dataset Hasil *Preprocessing* Data Untuk Prediksi dan Klasifikasi

Kabupaten Kota	Jenis Sekolah	Jumlah Sekolah	Jumlah Guru	Jumlah Murid	Jumlah Ruang Kelas
Kepulauan Seribu	Negeri	0,000	0,000	0,000	0,000
Kepulauan Seribu	Swasta	10,000	44,000	119,000	12,000
Jakarta Selatan	Negeri	7,000	43,000	371,000	20,000
Jakarta Selatan	Swasta	437,000	2500,000	19186,000	1448,000
Jakarta Timur	Negeri	4,000	24,000	187,000	8,000
Jakarta Timur	Swasta	0,000	2929,000	23368,000	1702,000
Jakarta Pusat	Negeri	608,000	30,000	238,000	11,000
Jakarta Pusat	Swasta	5,000	929,000	7135,000	523,000
Jakarta Barat	Negeri	208,000	0,000	0,000	0,000
Jakarta Barat	Swasta	435,000	2479,000	21138,000	1403,000
Jakarta Utara	Negeri	1,000	6,000	40,000	3,000
Jakarta Utara	Swasta	259,000	1464,000	11849,000	779,000

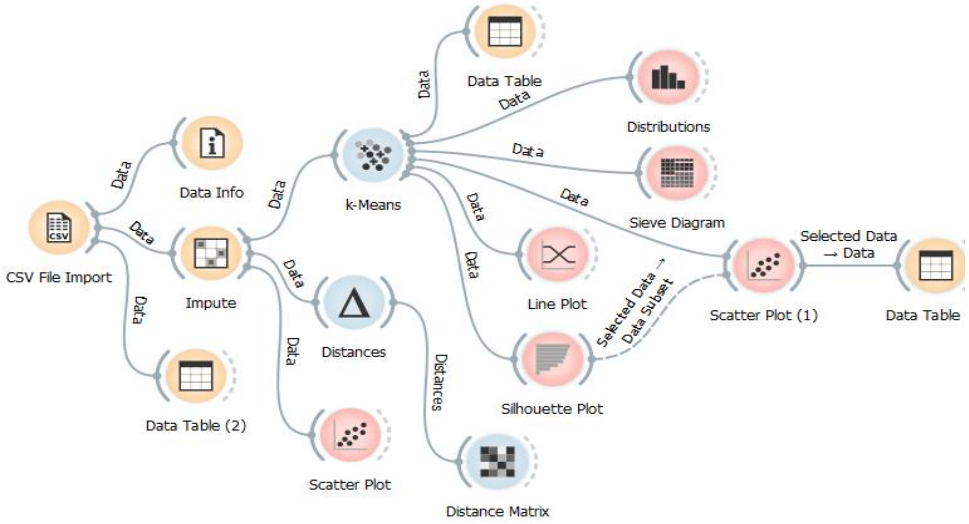
 DOI: https://doi.org/10.30998/faktorexacta.v15i2.13211



Gambar 2. *Plotting Data* yang Digunakan.

3.3 Hasil Transformation

Hasil pengkodean pada yang dipilih dari hasil *preprocessing* data dengan menggabungkan beberapa algoritma dalam satu proses transformasi, sehingga proses transformasi data yang dapat dilakukan untuk proses *data mining* seperti ditunjukan pada gambar 3.



Gambar 3. Proses Hasil Transformation/Pengkodean Data.

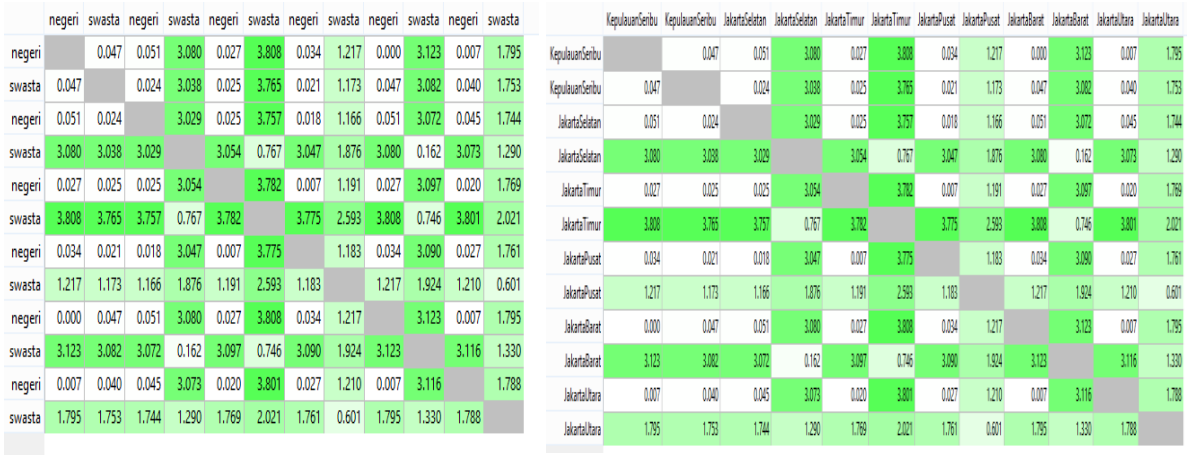
Pada proses ini data yang digunakan berformat csv. di *import* ke dalam kode program, kemudian pada bagian data yang telah di *import* dapat dilakukan *preprocessing* data dengan melakukan sedikit konfigurasi menggunakan teknik *impute* data berdasarkan *average* atau *most frequency* pada setiap atribut yang dibutuhkan. Jika data tersebut *cleaning* maka data dapat dilakukan proses data mining menggunakan beberapa algoritma diantaranya algoritma *Levenshtein Distance* dan *K-Means Clustering*.

3.4 Hasil *Data Mining*

Pada penelitian ini menggunakan 2 metode sebagai komparasi dari 2 algoritma dengan mencari *knowledge* baru untuk pengambilan keputusan.

Levenshtein Distance

Hasil menggunakan metode *Levenshtein Distance* dengan mensimulasikan data melalui model matriks pengukuran menggunakan jarak *Euclidean dan jumlah Rows* (baris) dari hasil *preprocessing* data seperti ditunjukan pada gambar 4. Metode *Levenshtein Distance* yang digunakan menggunakan *visual programming* dari bahasa pemograman *python*, dengan memadukan teknik pada *Distance* kemudian dikonversi ke dalam bentuk *Distance Matrix* yang ditunjukan pada gambar 3 di atas.



	negeri	swasta	negeri	swasta	negeri	swasta	negeri	swasta	negeri	swasta	negeri	swasta
negeri		0.047	0.051	3.080	0.027	3.808	0.034	1.217	0.000	3.123	0.007	1.795
swasta	0.047		0.024	3.038	0.025	3.765	0.021	1.173	0.047	3.082	0.040	1.753
negeri	0.051	0.024		3.029	0.025	3.757	0.018	1.166	0.051	3.072	0.045	1.744
swasta	3.080	3.038	3.029		3.054	0.767	3.047	1.876	3.080	0.162	3.073	1.290
negeri	0.027	0.025	0.025	3.054		3.782	0.007	1.191	0.027	3.097	0.020	1.769
swasta	3.808	3.765	3.757	0.767	3.782		3.775	2.593	3.808	0.746	3.801	2.021
negeri	0.034	0.021	0.018	3.047	0.007	3.775		1.183	0.034	3.090	0.027	1.761
swasta	1.217	1.173	1.166	1.876	1.191	2.593	1.183		1.217	1.924	1.210	0.601
negeri	0.000	0.047	0.051	3.080	0.027	3.808	0.034	1.217		3.123	0.007	1.795
swasta	3.123	3.082	3.072	0.162	3.097	0.746	3.090	1.924	3.123		3.116	1.330
negeri	0.007	0.040	0.045	3.073	0.020	3.801	0.027	1.210	0.007	3.116		1.788
swasta	1.795	1.753	1.744	1.290	1.769	2.021	1.761	0.601	1.795	1.330	1.788	

	KepulauanSeribu	KepulauanSeribu	JakartaSelatan	JakartaSelatan	JakartaTimur	JakartaTimur	JakartaPusat	JakartaPusat	JakartaBarat	JakartaBarat	JakartaUtara	JakartaUtara
KepulauanSeribu		0.047	0.051	3.080	0.027	3.808	0.034	1.217	0.000	3.123	0.007	1.795
KepulauanSeribu	0.047		0.024	3.038	0.025	3.765	0.021	1.173	0.047	3.082	0.040	1.753
JakartaSelatan	0.051	0.024		3.029	0.025	3.757	0.018	1.166	0.051	3.072	0.045	1.744
JakartaSelatan	3.080	3.038	3.029		3.054	0.767	3.047	1.876	3.080	0.162	3.073	1.290
JakartaTimur	0.027	0.025	0.025	3.054		3.782	0.007	1.191	0.027	3.097	0.020	1.769
JakartaTimur	3.808	3.765	3.757	0.767	3.782		3.775	2.593	3.808	0.746	3.801	2.021
JakartaPusat	0.034	0.021	0.018	3.047	0.007	3.775		1.183	0.034	3.090	0.027	1.761
JakartaPusat	1.217	1.173	1.166	1.876	1.191	2.593	1.183		1.217	1.924	1.210	0.601
JakartaBarat	0.000	0.047	0.051	3.080	0.027	3.808	0.034	1.217		3.123	0.007	1.795
JakartaBarat	3.123	3.082	3.072	0.162	3.097	0.746	3.090	1.924	3.123		3.116	1.330
JakartaUtara	0.007	0.040	0.045	3.073	0.020	3.801	0.027	1.210	0.007	3.116		1.788
JakartaUtara	1.795	1.753	1.744	1.290	1.769	2.021	1.761	0.601	1.795	1.330	1.788	

Bagian A Berdasarkan Jenis Sekolah Bagian B Berdasarkan Kabupaten/Kota

Gambar 4. Hasil *Data Mining* dengan Metode *Levenshtein Distance*

Simpulan bahwa pada gambar 4 bagian A diketahui bahwa string sumber pada baris "berdasarkan jenis sekolah" swasta dan negeri memiliki 12 (dua belas) karakter dan string target pada kolom "berdasarkan jenis sekolah" swasta dan negeri. Selanjutnya karakter ke-1 pada masing-masing string dibandingkan dan diketahui bahwa isi karakter ke-1 pada masing-masing string sama, maka nilai matriks yang diberikan sesuai dengan persamaan (4) yaitu $D(1,1)= D(1-1,1-1),sj=ti$. Jadi nilai matriks yang diberikan pada $D(1,1) = D(0,0)$ yang bernilai 0, kemudian nilai matriks yang diberikan pada $D(1,2)$ bernilai 0,024 karena persamaan (2) yaitu penyisipan dan seterusnya. Kemudian nilai matriks pada $D(1,1)$, $D(1,2)$ dan matriks lainnya diisi seperti yang digambarkan pada Gambar 4 bagian A. Teknik tersebut merupakan hasil simulasi data pengukuran menggunakan jarak *Euclidean dan jumlah Rows* (baris) berdasarkan jenis sekolah.

Perhitungan jarak *Levenshtein Distance* selanjutnya berjalan sampai semua nilai pada matriks terisi. Jarak *Levenshtein Distance* adalah nilai yang terdapat di bawah-kanan matriks, dan pada kasus string sumber "berdasarkan jenis sekolah" swasta dan negeri dan string target pada kolom "berdasarkan jenis sekolah" swasta dan negeri berada di $D(12,12)$. Setelah dilakukan seluruh perhitungan matriks diketahui hasil dari perhitungan jarak antara string sumber dan string target adalah bernilai 0 seperti yang digambarkan pada matriks gambar 4 bagian A.

Gambar 4 bagian B diketahui bahwa string sumber pada baris "berdasarkan kabupaten/kota" seperti kepulauan seribu, jakarta selatan, jakarta timur, jakarta pusat, jakarta barat dan jakarta utara yang memiliki 12 (dua belas) karakter dan string target pada kolom ke-6 kota "berdasarkan kabupaten/kota". Selanjutnya karakter ke-1 pada masing-masing string dibandingkan dan diketahui bahwa isi karakter ke-1 pada masing-masing string sama, maka nilai matriks yang diberikan sesuai dengan persamaan (4) yaitu $D(1,1)= D(1-1,1-1),sj=ti$. Jadi nilai matriks yang diberikan pada $D(1,1) = D(0,0)$ yang bernilai 0. Kemudian nilai matriks pada $D(1,1)$ diisi seperti yang digambarkan pada Gambar 4 bagian B dan seterusnya. Teknik tersebut merupakan hasil simulasi data pengukuran menggunakan jarak *Euclidean dan jumlah Rows* (baris) berdasarkan kabupaten/kota.

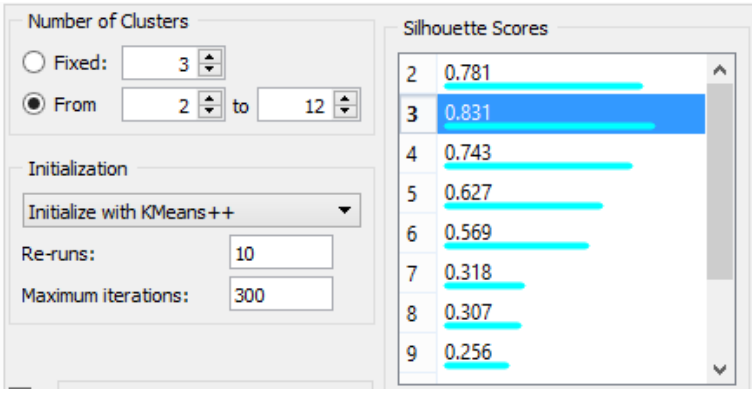
Pengukuran jarak *Levenshtein Distance* selanjutnya pada matriks bagian B berjalan sampai semua nilai pada matriks terisi. Jarak *Levenshtein Distance* adalah nilai yang terdapat di bawah-kanan matriks, dan pada kasus string sumber "berdasarkan kabupaten/kota" seperti kepulauan seribu, jakarta selatan, jakarta timur, jakarta pusat, jakarta barat dan jakarta utara dan string target pada kolom berdasarkan kabupaten/kota" seperti kepulauan seribu, jakarta selatan, jakarta timur, jakarta pusat, jakarta barat dan jakarta utara berada di $D(12,12)$. Setelah dilakukan seluruh perhitungan matriks diketahui hasil dari perhitungan jarak antara string sumber dan string target adalah $D(12,12) = D(0,0)$ yang bernilai 0 seperti yang digambarkan pada matriks gambar 4 bagian B.

Karena metode *Levenshtein Distance* mengukur kesamaan kedua string yakni string sumber dan string target, maka dapat disimpulkan kedua hasil *data mining* tersebut memiliki data yang *balance* serta memiliki korelasi yang sesuai dengan data setelah dilakukan *preposessing* data sebelumnya.

K-Means Clustering

Metode *K-Means Clustering* dalam pengelompokan data berdasarkan status sekolah dari hasil *preprocessing* data seperti ditunjukan pada gambar 5.

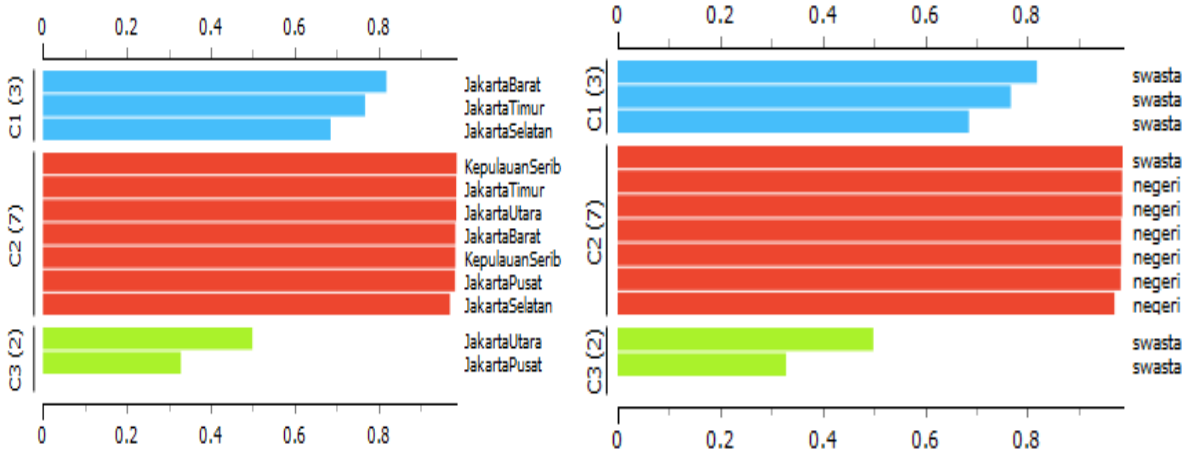
 DOI: https://doi.org/10.30998/faktorexacta.v15i2.13211



Gambar 5. Metode *K-Means Clustering Dengan Maksimum Iterasi.*

Gambar 5 merupakan konfigurasi algoritma *K-Means Clustering* yang digunakan pada penelitian ini dengan penomoran klaster mulai dari 2 sampai 12 data yang tersedia dari hasil *preprocessing* data, kemudian menghitung jarak penentuan titik pusat atau *centroid* terdekat menggunakan nilai maksimum iterasi 10 sampai 300 dengan cara mengambil data secara terpusat pada inisialisasi *K-Means++*.

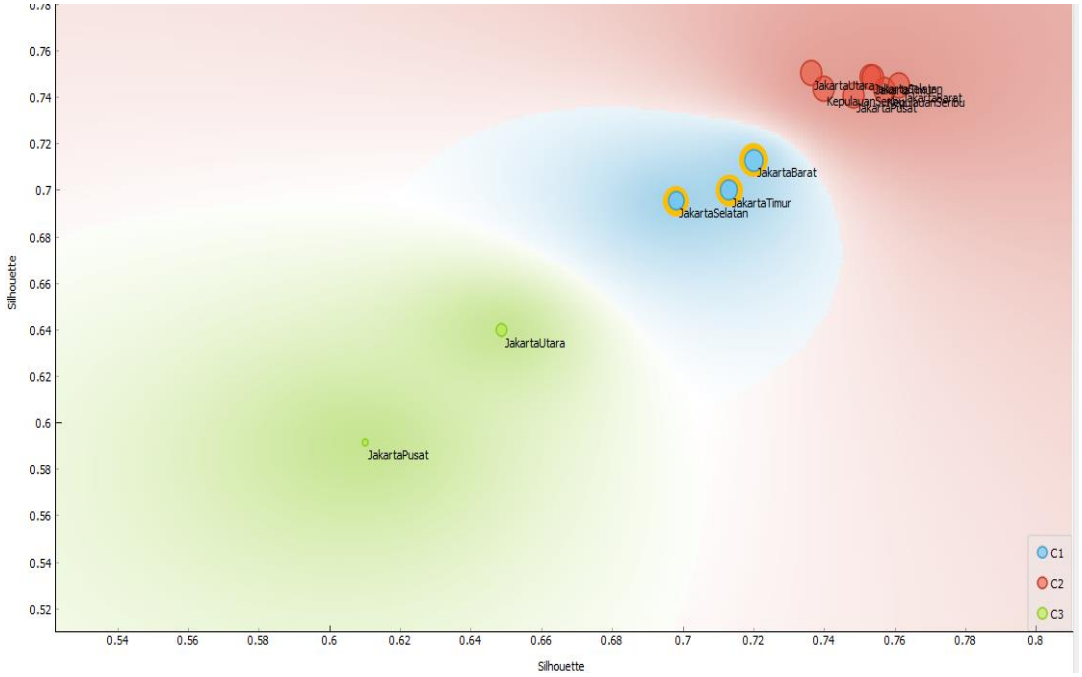
Berdasarkan *Average* atau *most frequency* dari 4 atribut yaitu jumlah sekolah, jumlah guru, jumlah murid, jumlah ruang kelas, maka *Silhouette Plot* simulasi data pengukuran menggunakan jarak *Euclidean* secara grup klaster ditunjukan pada gambar 6.



Bagian A Berdasarkan Kabupaten/Kota Bagian B Berdasarkan Jenis Sekolah

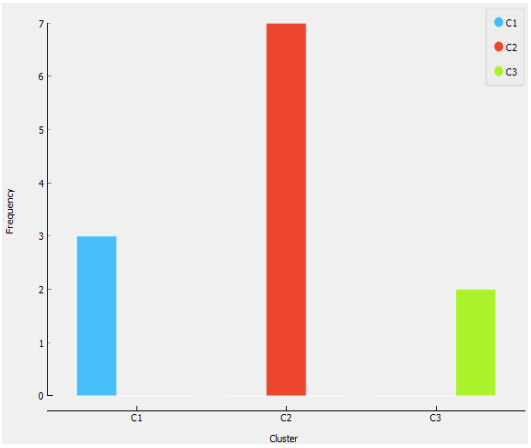
Gambar 6. Hasil *Silhouette Plot* Menggunakan Jarak *Euclidean.*

Jumlah *cluster* pada penelitian ini sebanyak 3 *cluster*, untuk ke-3 *cluster* ini akan menentukan hasil penelitian mengenai sekolah mana yang termasuk kategori terbanyak, sedikit dan paling sedikit dari 4 atribut yaitu jumlah sekolah, jumlah guru, jumlah murid, jumlah ruang kelas berdasarkan kabupaten/kota dan jenis sekolah. Gambar 6 bagian A termasuk C1 sebanyak 3 sekolah berdasarkan kabupaten/kota yaitu, Jakarta Barat, Jakarta Timur dan Jakarta Selatan. C2 sebanyak 7 sekolah berdasarkan kabupaten/kota yaitu, Kepulauan Seribu, Jakarta Timur, Jakarta Utara, Jakarta Barat, Kepulauan Seribu, Jakarta Pusat dan Jakarta Selatan. C3 sebanyak 2 sekolah berdasarkan kabupaten/kota yaitu, Jakarta Utara dan Jakarta Pusat. Gambar 6 bagian B termasuk C1 sebanyak 3 sekolah berdasarkan jenis sekolah yaitu swasta, swasta dan swasta. C2 sebanyak 7 sekolah berdasarkan jenis sekolah yaitu swasta, negeri, negeri, negeri, negeri, negeri dan negeri. C3 sebanyak 2 sekolah berdasarkan jenis sekolah yaitu swasta dan swasta.

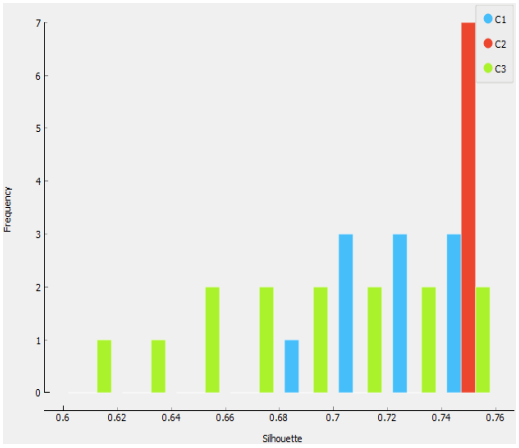


Gambar 7. Hasil *Plotting Silhouette Score* Berdasarkan Kabupaten/Kota

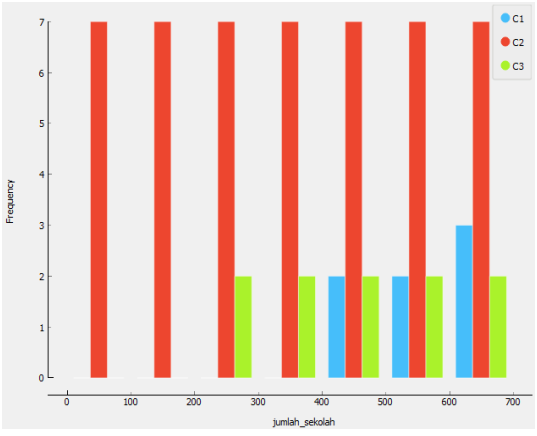
Gambar 8 merupakan hasil kumulatif data yang terdistribusi berdasarkan frekuensi jumlah data pada *cluster*, *plotting silhouette score*, jumlah sekolah, jumlah guru, jumlah murid, jumlah ruang kelas Sekolah TK berdasarkan status dan kabupaten/kota administrasi provinsi DKI Jakarta.



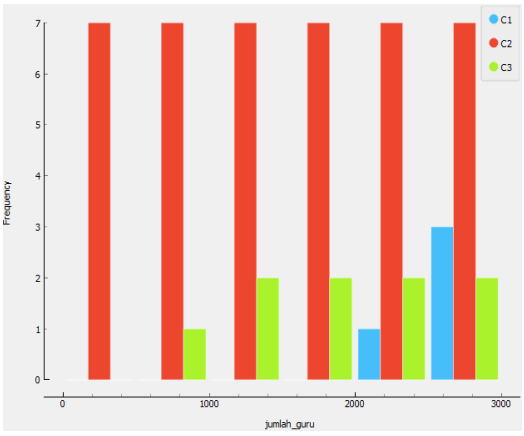
Berdasarkan Jumlah *Cluster*



Berdasarkan *silhouette score*

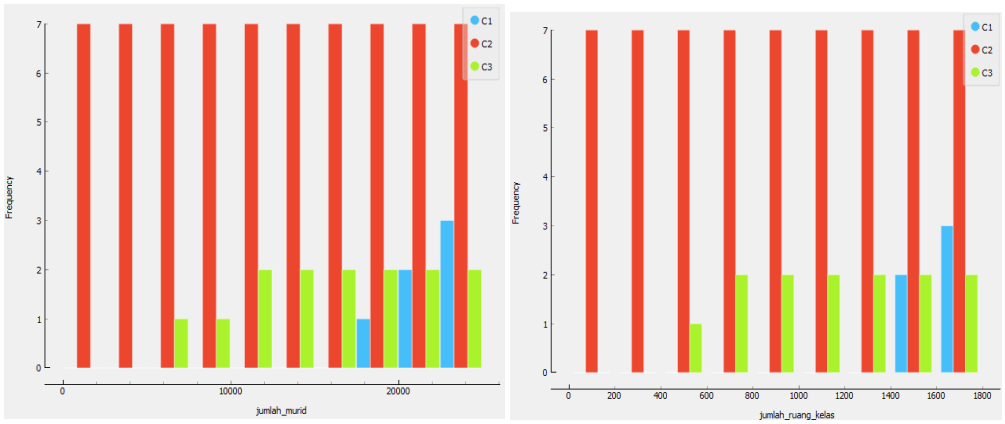


Berdasarkan Jumlah Sekolah



Berdasarkan Jumlah Guru

 DOI: https://doi.org/10.30998/faktorexacta.v15i2.13211



Berdasarkan Jumla Murid Berdasarkan Jumlah Ruang Kelas

Gambar 8. Hasil *Plotting* Kumulatif Data Terdistribusi

Hasil *plotting* komulatif data pada gambar 8 menjelaskan bahwa jumlah *cluster* terbanyak adalah *cluster* C2 yakni 7 sekolah TK berdasarkan kabupaten/kota dari C1 dan C3, kemudian berdasarkan *silhouette score* kurang dari 0,76 dari 12 data sekolah TK berdasarkan kabupaten/kota dengan persentase 100% maka dapat diasumsikan *cluster* C1 dengan persentase 25%, *cluster* C2 dengan persentase 58,33% dan *cluster* C3 dengan persentase 16,67%. Pada kumulatif berdasarkan jumlah sekolah kurang dari 700 dari 12 data sekolah TK berdasarkan kabupaten/kota dengan persentase 100% maka dapat diasumsikan *cluster* C1 dengan persentase 25%, cluster C2 dengan persentase 58,33% dan cluster C3 dengan persentase 16,67%, kemudian kumulatif berdasarkan jumlah guru kurang dari 3000 dari 12 data sekolah TK berdasarkan kabupaten/kota dengan persentase 100% maka dapat diasumsikan *cluster* C1 dengan persentase 25%, cluster C2 dengan persentase 58,33% dan cluster C3 dengan persentase 16,67%, kumulatif berdasarkan jumlah murid kurang dari 25000 dari 12 data sekolah TK berdasarkan kabupaten/kota dengan persentase 100% maka dapat diasumsikan *cluster* C1 dengan persentase 25%, cluster C2 dengan persentase 58,33% dan cluster C3 dengan persentase 16,67%, kumulatif berdasarkan jumlah ruang kelas kurang dari 1800 dari 12 data sekolah TK berdasarkan kabupaten/kota dengan persentase 100% maka dapat diasumsikan *cluster* C1 dengan persentase 25%, cluster C2 dengan persentase 58,33% dan cluster C3 dengan persentase 16,67%. Maka dapat disimpulkan bahwa presentase *cluster* terbesar adalah *cluster* C2 dengan persentase 58,33%.

Interpretation / Evaluation

Tabel 3 di bawah ini merupakan hasil *plotting silhouette score* pada *cluster* C1 yaitu Jakarta Selatan, Jakarta Timur dan Jakarta Barat. Maka dapat disimpulkan jakarta selatan masuk ke dalam group G1 dengan jenis sekolah swasta yang memiliki *silhouette score* 0,691335, jakarta timur masuk ke dalam group G1 dengan jenis sekolah swasta yang memiliki *silhouette score* 0,708418, jakarta barat masuk ke dalam group G1 dengan jenis sekolah swasta yang memiliki *silhouette score* 0,718406.

Tabel 3. *Silhouette Score Cluster* C1

Group	Kabupaten Kota	Jenis Sekolah	*Cluster*	*Silhouette*	Jumlah Sekolah	Jumlah Guru	Jumlah Murid	Jumlah Ruang Kelas
G1	Jakarta Selatan	Swasta	C1	0,691335	437,000	2500,000	19186,000	1448,000
G1	Jakarta Timur	Swasta	C1	0,708418	608,000	2929,000	23368,000	1702,000
G1	Jakarta Barat	Swasta	C1	0,718406	435,000	2479,000	21138,000	1403,000

Pada tabel 4 di bawah ini merupakan hasil *plotting silhouette score* pada *cluster* C2 yaitu Kepulauan Seribu, Kepulauan Seribu, Jakarta Selatan, Jakarta Timur, Jakarta Pusat, Jakarta Barat, Jakarta Utara . Maka dapat disimpulkan kepulauan seribu masuk ke dalam group G1 dengan jenis sekolah negeri yang memiliki *silhouette score* 0,747291, kepulauan seribu masuk ke dalam group G1 dengan jenis sekolah swasta yang memiliki *silhouette score* 0,747778, jakarta selatan masuk ke dalam group G1 dengan jenis sekolah negeri yang memiliki *silhouette score* 0,745171, jakarta timur masuk ke dalam group G1 dengan jenis sekolah negeri yang memiliki *silhouette score* 0,747631, jakarta pusat masuk ke dalam group G1 dengan jenis sekolah negeri yang memiliki *silhouette score* 0,747181, jakarta barat masuk ke dalam group G1 dengan jenis sekolah negeri yang memiliki *silhouette*

score 0,747291 dan jakarta utara masuk ke dalam group G1 dengan jenis sekolah negeri yang memiliki *silhouette score* 0,747619.

Tabel 4. *Silhouette Score Cluster* C2

Group	Kabupaten Kota	Jenis Sekolah	*Cluster*	*Silhouette*	Jumlah Sekolah	Jumlah Guru	Jumlah Murid	Jumlah Ruang Kelas
G1	Kepulauan Seribu	Negeri	C2	0,747291	0,000	0,000	0,000	0,000
G1	Kepulauan Seribu	Swasta	C2	0,747778	10,000	44,000	119,000	12,000
G1	Jakarta Selatan	Negeri	C2	0,745171	7,000	43,000	371,000	20,000
G1	Jakarta Timur	Negeri	C2	0,747631	4,000	24,000	187,000	8,000
G1	Jakarta Pusat	Negeri	C2	0,747181	5,000	30,000	238,000	11,000
G1	Jakarta Barat	Negeri	C2	0,747291	0,000	0,000	0,000	0,000
G1	Jakarta Utara	Negeri	C2	0,747619	1,000	6,000	40,000	3,000

Pada tabel 5 di bawah ini merupakan hasil *plotting silhouette score* pada *cluster* C3 yaitu Jakarta Pusat dan Jakarta Utara. Maka dapat disimpulkan Jakarta Pusat masuk ke dalam group G1 dengan jenis sekolah swasta yang memiliki *silhouette score* 0,601115 dan Jakarta Utara masuk ke dalam group G1 dengan jenis sekolah swasta yang memiliki *silhouette score* 0,647377.

Tabel 5. *Silhouette Score Cluster* C3

Group	Kabupaten Kota	Jenis Sekolah	*Cluster*	*Silhouette*	Jumlah Sekolah	Jumlah Guru	Jumlah Murid	Jumlah Ruang Kelas
G1	Jakarta Pusat	Swasta	C3	0,601115	208,000	929,000	7135,000	523,000
G1	Jakarta Utara	Swasta	C3	0,647377	259,000	1464,000	11849,000	779,000

4. PENUTUP

4.1 Simpulan

Berdasarkan hasil penelitian yang telah dilakukan dengan memodelkan data status sekolah TK kabupaten dan kota administrasi provinsi DKI Jakarta yang diambil dari data sekunder pada portal data.jakarta.go.id. serta menggabungkan desain pemodelan *machine learning*, didapatkan hipotesis dugaan bahwa metode klasifikasi pemodelan dengan menerapkan 2 algoritma yang digunakan *Levenshtein Distance dan K-Means Clustering* dapat digunakan secara efektif untuk melakukan klasifikasi dan *forecasting* sesuai dengan penentuan jumlah klaster sekolah yang termasuk ke dalam kategori terbanyak, sedikit dan paling sedikit berdasarkan 4 atribut pada tiap sekolah yaitu jumlah sekolah, jumlah guru, jumlah murid, jumlah ruang kelas berdasarkan kabupaten/kota dan jenis sekolah. Dilihat dari hasil plot kumulatif bahwa jumlah *cluster* terbanyak adalah *cluster* C2 yakni 7 sekolah TK berdasarkan kabupaten/kota dari C1 dan C3, kemudian berdasarkan *silhouette score* kurang dari 0,76 dari 12 data sekolah TK berdasarkan kabupaten/kota dengan persentase 100%, maka dapat diasumsikan *cluster* C1 dengan persentase 25%, *cluster* C2 dengan persentase 58,33% dan *cluster* C3 dengan persentase 16,67%. Sehingga dari hasil kumulatif jumlah klaster tersebut pemprov DKI selaku penyelenggara pendidikan dan calon peserta didik dapat menentukan sekolah-sekolah TK yang perlu dipilih dan diklasifikasikan baik dari sisi penambahan jumlah sekolah pada regional tertentu yang minim akan jumlah sekolah TK, penambahan jumlah guru bagi sekolah yang minim guru, penambahan jumlah murid bagi sekolah yang membutuhkan kouta murid sedikit berdasarkan regional tertentu serta jumlah ruang kelas pada sekolah yang minim akan ruang kelas untuk melaksanakan pertemuan tatap muka dalam kegiatan belajar mengajar.

4.2. Saran

Penelitian ini diharapkan dapat diaplikasikan dan membantu dalam melakukan klasifikasi penentuan sekolah-sekolah sebagai wujud kepedulian pemerintah terhadap pendidikan sesuai dengan visi misi dinas pendidikan DKI Jakarta untuk mewujudkan pendidikan yang tuntas dan berkualitas. Hasil penelitian ini belum dikatakan sempurna sehingga perlunya ada pengembangan berkelanjutan dalam pemodelan data dengan kajian dan komparasi pada algoritma yang lebih efektif.

DAFTAR PUSTAKA

[1] S. J. Russell and P. Norvig, *Artificial Intelligence (A Modern Approach ThirdEdition Stuart)*, vol. 48. 2016.

[2] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*, vol. 9781107057. 2013.

[3] R. Verganti, L. Vendraminelli, and M. Iansiti, "Innovation and Design in the Age of Artificial Intelligence," *J. Prod. Innov. Manag.*, vol. 37, no. 3, pp. 212–227, 2020.

[4] Y. Jin, H. Wang, and C. Sun, "Introduction to Machine Learning," in *Studies in Computational Intelligence*, vol. 975, 2021, pp. 103–145.

[5] S. Vieira, W. H. Lopez Pinaya, and A. Mechelli, "Introduction to machine learning," in *Machine Learning: Methods and Applications to Brain Disorders*, 2019, pp. 1–20.

[6] O. Pentakalos, "Introduction to machine learning," in *CMG IMPACT 2019*, 2019.

[7] N. A. Parasa, J. V. Namgiri, S. N. Mohanty, and J. K. Dash, "Introduction to Unsupervised Learning in Bioinformatics," in *Data Analytics in Bioinformatics*, 2021, pp. 35–49.

[8] A. Ng, *Clustering - Unsupervised Learning Introduction*. 2021, pp. 209–218.

[9] S. Ullman, T. Poggio, D. Harari, D. Zysman, and D. Seibert, "9.54 Class 13 Unsupervised learning Clustering," *Unsupervised Learn. Slides*, p. 54, 2014.

[10] G. Hicham, Y. Abdallah, and B. Mostapha, "Introduction of the weight edition errors in the Levenshtein distance," *Int. J. Adv. Res. Artif. Intell.*, vol. 1, no. 5, 2012.

[11] F. Indriyani and E. Irfiani, "Clustering Data Penjualan pada Toko Perlengkapan Outdoor Menggunakan Metode K-Means (Clustering Sales Data at Outdoor Equipment Stores Using KMeans Method)," *JUITA J. Inform.*, vol. 7, no. 2, pp. 109–113, 2019.

[12] M. Chau, R. Cheng, B. Kao, and J. Ng, "Uncertain data mining: An example in clustering location data," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2006, vol. 3918 LNAI, pp. 199–204.

[13] K. Fatmawati and A. P. Windarto, "DATA MINING: PENERAPAN RAPIDMINER DENGAN K-MEANS CLUSTER PADA DAERAH TERJANGKIT DEMAM BERDARAH DENGUE (DBD) BERDASARKAN PROVINSI," *Comput. Eng. Sci. Syst. J.*, vol. 3, no. 2, p. 173, 2018.

[14] P. Novianti, D. Setyorini, and U. Rafflesia, "K-means cluster analysis in earthquake epicenter clustering," *Int. J. Adv. Intell. Informatics*, vol. 3, no. 2, pp. 81–89, 2017.

[15] M. H. Adiya and Y. Desnelita, "Jurnal Nasional Teknologi dan Sistem Informasi Penerapan Algoritma K-Means Untuk Clustering Data Obat-Obatan Pada RSUD Pekanbaru," *Nas. Teknol. dan Sist. Inf.*, vol. 01, pp. 17–24, 2019.

[16] S. Ghosal, R. Bhattacharyya, and M. Majumder, "Impact of complete lockdown on total infection and death rates: A hierarchical cluster analysis," *Diabetes Metab. Syndr. Clin. Res. Rev.*, vol. 14, no. 4, pp. 707–711, 2020.

[17] A. Nastuti and S. Z. Harahap, "TEKNIK DATA MINING UNTUK PENENTUAN PAKET HEMAT SEMBAKO DAN KEBUTUHAN HARIAN DENGAN MENGGUNAKAN ALGORITMA FP-GROWTH (STUDI KASUS DI ULFAMART LUBUK ALUNG)," *J. Inform.*, vol. 7, no. 3, pp. 111–119, 2019.

[18] N. Nidheesh, K. A. Abdul Nazeer, and P. M. Ameer, "An enhanced deterministic K-Means clustering algorithm for cancer subtype prediction from gene expression data," *Comput. Biol. Med.*, vol. 91, pp. 213–221, 2017.