BAB II

TINJAUAN PUSTAKA DAN DASAR TEORI

2.1. Tinjauan Pustaka

Penelitian ini mengacu pada literatur yang bersumber dari jurnal-jurnal terkait, yang digunakan sebagai referensi dan panduan dalam penyusunan penelitian ini. Sumber-sumber pustaka yang dimaksud mencakup:

Syaputri, D., Noprita, P. H., & Romelah, S. (2021) melakukan penelitian dengan menggunakan metode *clustering K-means*. Penelitian ini bertujuan untuk mengelompokkan distribusi sosial ekonomi di masyarakat perumahan kota Pekanbaru berdasarkan faktor demografi penduduk. Selain itu, penelitian ini juga bertujuan untuk menggali pola hubungan sosial ekonomi di antara warga perumahan Pekanbaru berdasarkan demografi mereka. Data yang digunakan adalah data hasil dari sebaran kuisioner kepada 300 penduduk di Kota Pekanbaru. Hasil dari analisis menggunakan metode *K-means* menghasilkan 4 kluster yang dianggap optimal, dengan nilai indeks Davies-Bouldin (DBI) sebesar 0,87.

Di sisi lain, Puspasari, et al. (2021) menjalankan penelitian dengan fokus yang berbeda, yaitu mengelompokkan pelamar dan lowongan kerja berdasarkan keterampilan, gaji, lokasi, dan faktor lainnya. Tujuan utamanya adalah memberikan rekomendasi pekerjaan yang lebih personal dan sesuai dengan profil pelamar, dengan mempertimbangkan tingkat kesesuaian antara data pelamar dan kebutuhan perusahaan. Penelitian ini bertujuan meningkatkan efisiensi dalam pencocokan pekerjaan, khususnya untuk pelamar dengan gelar sarjana. Hasilnya berupa sistem yang memiliki fungsi rekomendasi yang mengelompokkan pelamar dan

lowonganpekerjaan berdasarkan keterampilan utama, gaji, lokasi, dan keterampilan lainnya dengan memberikan bobot pada variabel-variabel tersebut dan sistem berjalan dengan baik dengan tingkat kepuasan pengguna sebesar 87.6%.

Meskipun tujuan penelitian keduanya berbeda, keduanya menggunakan metode K-means *clustering* sebagai alat analisis untuk mencapai tujuan mereka. Penelitian Syaputri, et al. (2021) berfokus pada aspek sosial ekonomi berdasarkan demografi, sementara Puspasari, et al. (2021) lebih menekankan pada pencocokan pekerjaan untuk pelamar.

Amoozad Mahdiraji, H., et al. (2022) menganalisis data pelanggan menggunakan model RFM yang dimodifikasi dan K-means untuk mengidentifikasi pelanggan terbaik. Mereka menentukan jumlah *cluster* dengan analisis dua langkah K-means menggunakan indeks *Silhouette, Davies-Bouldin,* dan *Calinski-Harabasz*. Kemudian, mereka menggunakan metode *Best-Worst Method* (*BWM*) dan Total *Area based on Vector Orthogonal (TAOV)* untuk mengurutkan *cluster*. Selanjutnya, mereka menggunakan aturan asosiatif dan algoritma Apriori untuk mengungkap pola perilaku pelanggan.

Hasil penelitian ini mengelompokkan pelanggan menjadi 6 *cluster* berdasarkan analisis terhadap 20.000 data yaitu nasabah yang sangat loyal, loyal, berinteraksi tinggi, berinteraksi sedang, berinteraksi rendah, dan nasabah yang hilang. Kemudian mengidentifikasi pola perilaku keuangan nasabah berdasarkan demografi dan transaksi, serta mengklasifikasikan jenis pelanggan. Ini membantu merumuskan strategi interaksi yang sesuai untuk setiap jenis pelanggan.

Martín, I., Mariello, A., Battiti, R., & Hernández, J. A. (2018) memfokuskan pada perumusan masalah prediksi gaji sebagai tugas klasifikasi, di mana pekerjaan diklasifikasikan ke dalam empat kelas yang mewakili rentang gaji. Dengan pendekatan ini yang memisahkan rentang gaji, bukan nilai gaji yang kontinu, diharapkan dapat meningkatkan akurasi prediksi. Studi ini melakukan perbandingan berbagai pengklasifikasi, termasuk linear models, logistic regression, *K-nearest neighbors* (KNN), *multi-layer perceptrons* (MLP), *support vector machines* (SVM*), random forests, decision trees,* dan *ensembles of these models*, untuk mengidentifikasi model yang memiliki akurasi terbaik dalam memprediksi kategori rentang gaji. Didapatkan akurasi terbaik dari metode decision tree yaitu 84%. Dalam penelitian ini juga digunakan metode K-means *clustering* untuk mempartisi data menjadi kelompok-kelompok yang homogen.

Sejalan dengan penelitian Martín, I., at. al (2018) yang memfokuskan pada prediksi gaji dengan pendekatan klasifikasi rentang gaji, penelitian Sayan, D, et. al (2020) juga memiliki fokus serupa dalam memprediksi gaji. Namun, penelitian ini menggunakan teknik regresi untuk membuat grafik dan memprediksi gaji. Teknik regresi yang diterapkan melibatkan regresi linier dan regresi polinomial. Tujuan utama dari penelitian Sayan, D et. al. (2020) adalah memberikan bantuan kepada karyawan dalam menentukan gaji berdasarkan kualifikasi mereka serta memantau pertumbuhan karier mereka dalam suatu bidang. Hasilnya adalah aplikasi yang memprediksi gaji berdasarkan database gaji organisasi dengan menggunakan algoritma prediksi.

Sementara itu, Munti, NYS, Nurcahyo, GW, & Santony, J. (2018) menggunakan algoritma K-means *clustering* dalam menentukan gaji karyawan tetap dan karyawan kontrak (Studi Kasus Di Pt Indomex Dwijaya Lestari).” Hasil penelitian ini menunjukkan bahwa penerapan algoritma K-means *clustering* dalam menganalisis gaji karyawan di PT Indomex Dwijaya Lestari mampu mengelompokkan gaji secara efektif ke dalam *cluster* yang berbeda berdasarkan faktor internal yang mempengaruhi kinerja karyawan. Dalam penelitian ini, digunakan 60 data sampel dan dilakukan 4 kali iterasi hinga nilai rata-rata *cluster* tidak berubah dan terdapat 2 *cluster* yang dihasilkan. *cluster* pertama adalah *cluster* dengan gaji rendah dimana terdapat 38 karyawan, sedangkan *cluster* kedua adalah *cluster* dengan gaji tinggi sebanyak 22 karyawan.

Dari beberapa penelitian yang telah membahas berbagai aspek penggunaan metode K-means *clustering* dalam konteks yang berbeda, penelitian ini memiliki fokus yang unik yaitu pada pengelompokan gaji. Penelitian ini akan mengeksplorasi potensi metode K-means *clustering* dalam mengelompokkan gaji karyawan berdasarkan karakteristik demografis mereka.

Tabel 2. 1 Penelitian Terdahulu

No	Peneliti	Objek	Masalah	Metode	Hasil
1.	Syaputri, D., Noprita, P. H., & Romelah, S. (2021)	Masyarakat Perumahan di Kota Pekanbaru	Pengelompokan Distribusi Sosial Ekonomi	Algoritma *K-means*	percobaan dengan jumlah *Cluster* = 4 merupakan komponen percobaan terbaik dengan formulasi pada *Cluster* 2 dengan nilai DBI sebesar 0,87.
2.	Puspasari, B. D., et al. (2021)	Sekolah Tinggi Teknik Atlas Nusantara	Metode efektif pengelompokan dalam sistem	*K-means*	Sistem yang memiliki fungsi rekomendasi yang mengelompokkan

		(STTAR) Malang	rekomendasi pekerjaan		pelamar dan lowongan pekerjaan
3.	Amoozad Mahdiraji, H., et al. (2022)	Nasabah di industri perbankan	Pengelompokan data pelanggan dan identifikasi pola perilaku nasabah.	RFM dan *K-means*	Terdapat 6 *cluster* yaitu nasabah yang sangat loyal, loyal, berinteraksi tinggi, berinteraksi sedang, berinteraksi rendah, dan nasabah yang hilang
4.	Martín, I., Mariello, A., Battiti, R., & Hernández, J. A. (2018)	Gaji, studi kasus di Spanyol	Prediksi gaji dengan pendekatan klasifikasi	*Linear models, logistic regression,(* KNN*), (*MLP*), (*SVM*), random forests, decision trees, ensembles of these models* dan *K-means*	*K-means* menghasilkan 9 kluster yang memisahkan penawaran berdasarkan rentang gaji dan *decision trees* mendapat akurasi terbaik 84%
5.	Sayan, D, et. al (2020)	Gaji	Prediksi gaji dengan pendekatan klasifikasi	Regresi linier dan regresi polinomial	Aplikasi komputerisasi yang bertujuan untuk memprediksi gaji
6.	Munti, N.Y.S, Nurcahyo, G.W, & Santony, J. (2018)	Pt Indomex Dwijaya Lestari	Menentukan Gaji Karyawan Tetap Dan Karyawan Kontrak	*K-means Clustering*	Terdapat 2 *Cluster* yang dihasilkan. C1 : *Cluster* dengan gaji rendah terdapat 38 karyawan, sedangkan C2: *Cluster* dengan gaji tinggi sebanyak 22 karyawan.
7.	Yanti, N (2023)	Gaji	*Clustering* gaji	*K-means Clustering*	Pembentukan *Cluster* optimal berdasarkan kesamaan karakteristik demografi dan mengukur hasil *cluster*ing berdasarkan metrik *Davies-Bouldin* dan *Silhouette Score*.

2.2. Dasar Teori

2.2.1. Gaji

Menurut Rybak (2019), gaji merupakan insentif yang sangat efektif dalam meningkatkan produktivitas, tetapi cara pengaturan dan pembentukannya masih

belum sepenuhnya terperinci. Sesuai dengan hukum di Ukraina, gaji diartikan sebagai penggantian yang biasanya berupa uang yang diberikan oleh majikan kepada pekerja sesuai dengan kontrak kerja mereka sebagai imbalan atas pekerjaan yang telah dilakukan. Besarnya gaji dapat bervariasi tergantung pada sejumlah faktor, termasuk kompleksitas pekerjaan, kualifikasi profesional, kebijakan perusahaan, serta hasil kerja dan kinerja ekonomi perusahaan.

Salah satu komponen penting dalam peraturan di negara adalah pengakuan terhadap Peraturan Upah Minimum. Di negara-negara maju, upah minimum berfungsi untuk menjaga tingkat penghasilan minimum yang diperlukan untuk kehidupan layak dan ditentukan melalui salah satu dari dua pendekatan berikut (Rybak 2019):

a. Sebagai persentase dari rata-rata upah;
b. Berdasarkan pendapatan yang diterima oleh rata-rata anggota keluarga.

Dalam kedua kasus tersebut, besarnya upah minimum berhubungan dengan kondisi ekonomi umum, situasi produksi, dan faktor-faktor yang memengaruhi pendapatan para pekerja dalam masyarakat.

2.2.2. Demografi Profil

Karakteristik demografis merujuk pada atribut-atribut individu, seperti umur, gender, tingkat pendidikan, penghasilan, dan asal negara, yang sering diaplikasikan dalam penelitian untuk mengidentifikasi pandangan dan persepsi individu (Papastathopoulos, A., et.al 2020).

Salah satu aspek demografis, yaitu jenis kelamin, dapat berdampak pada cara penduduk menginterpretasikan pandangan mereka (Long dan Kayat 2011;

Nunkoo dan Gursoy 2012 dalam Papastathopoulos, A., et.al (2020). Dalam sebuah studi oleh Thrane (2008) dalam Papastathopoulos, A., et.al (2020), disimpulkan bahwa pria yang bekerja di industri pariwisata mendapatkan gaji sekitar 20% lebih tinggi setiap tahunnya dibandingkan dengan rekan-rekan wanita mereka.

Tingkat pendidikan telah dimanfaatkan dalam berbagai penelitian dengan tujuan yang beragam. Pendidikan sering digunakan sebagai indikator status sosial (Ghubash 1992 dalam Papastathopoulos, A., et.al (2020) dan sebagai petunjuk kemampuan individu untuk mendapatkan pekerjaan dengan upah tinggi, yang menawarkan kesempatan bagi mereka untuk meningkatkan kualitas hidup (Thrane 2008 dalam Papastathopoulos, A., et.al (2020).

Kewarganegaraan adalah istilah yang mengacu pada kepemilikan dokumen-dokumen resmi yang diterbitkan oleh pemerintah suatu negara. Ini merupakan salah satu faktor demografis yang paling signifikan dalam memengaruhi pandangan individu (Prayag dan Ryan 2012 dalam Papastathopoulos, A., et.al (2020). Seringkali, penduduk merasa memiliki ikatan kuat dengan negara tempat mereka menjadi warganegara dan mengidentifikasi diri mereka dengan kebangsaan mereka. Dalam literatur pariwisata, mereka juga kerap disebut sebagai warga lokal, penduduk asli, atau orang-orang pribumi (Rahman 2010; Hernández dan Mercader 2015 dalam Papastathopoulos, A., et.al (2020).

2.2.3. Algoritma K-means *Clustering*

Menurut K´aroly et.al dalam Ilbeigipour, S., et.al (2022) algoritma K-means menggunakan parameter k yang diambil dari input untuk membagi n sampel menjadi k *cluster* dengan tujuan meningkatkan kesamaan internal di dalam setiap

cluster sementara menjaga tingkat kemiripan eksternal *cluster* serendah mungkin. Tingkat kemiripan dalam setiap *cluster* diukur dengan menghitung rata-rata jarak antara sampel-sampel dalam *cluster* tersebut. Pada dasarnya, algoritma ini merupakan metode heuristik yang bertujuan untuk mengurangi kesalahan kuadrat yang dijelaskan dalam Persamaan 2.1.

$$E = \sum_{i=0}^{k} \sum_{p \varepsilon Ci} |p - mi|2 \qquad \text{.....(2.1)}$$

Dalam konteks ini, E mewakili total kesalahan kuadrat dari seluruh sampel data, p adalah representasi dari data sampel yang termasuk dalam *cluster* Ci, dan mi adalah nilai rata-rata dari semua sampel dalam *cluster* Ci.

K-means clustering merupakan salah satu teknik dalam *unsupervised learning* yang berfokus pada analisis data tanpa ada informasi label sebelumnya(Liao, Q. Z., et. Al., 2022). Teknik ini digunakan untuk mengelompokkan data ke dalam kelompok-kelompok yang serupa dan menentukan pusat-pusat kelompok tersebut. Untuk menggunakan metode ini, peneliti harus menentukan jumlah kelompok yang diinginkan terlebih dahulu. Proses *K-means* kemudian berulang kali melakukan perubahan posisi pusat-pusat kelompok untuk mengurangi sebanyak mungkin variasi atau perbedaan dalam setiap kelompok. Dalam konteks ini, tujuan utamanya adalah meminimalkan perbedaan rata-rata antara data dalam satu kelompok dengan pusat kelompok yang didefinisikan oleh titik-titik data dalam kelompok tersebut (Hastie et al., 2009 dalam Liao, Q. Z., et. Al., 2022).

Zhang, G., Zhang, C., & Zhang, H. (2018) menyatakan bahwa algoritma *K-means* memiliki karakteristik utama sebagai berikut:

a. Kemampuan untuk menangani dengan efisien sejumlah besar data.
b. Selalu berhenti pada titik optimal lokal.
c. Hanya beroperasi pada nilai-nilai numerik.
d. Kelompok-kelompok yang dihasilkan cenderung memiliki bentuk yang cembung.

2.2.4. Jarak Eucledian

Pengukuran jarak yang diterapkan adalah jarak *euclidean.* Jarak *euclidean* merupakan salah satu metode untuk mengukur jarak yang digunakan untuk menghitung jarak antara dua titik data dalam ruang *Euclidean*. Ruang *Euclidean* dapat mencakup dua dimensi, tiga dimensi, atau bahkan lebih banyak dimensi (Faisal, M., et. al., 2020). Untuk mengevaluasi tingkat kemiripan data dengan rumus *Euclidean* Distance dapat diterapkan dengan persamaan 2.2 (H. Anton. 1993 dalam Faisal 2020):

$$d_{ij} = \sqrt{\sum_{k=1}^{n}(Xik - Yjk)^2} \qquad \text{.....(2.2)}$$

Dalam konteks ini, "d" merupakan jarak antara titik data i dan j, di mana i adalah pusat dari *cluster* data dan j adalah salah satu data dalam *cluster* tersebut. "k" adalah simbol yang merepresentasikan setiap data, "n" adalah jumlah total data, xik adalah data pada pusat *cluster* ke-k, dan yjk adalah data pada masing-masing data dalam *cluster* ke-k.

2.2.5. *Clustering*

Jurczyk, T. (2021) mendefinisikan *clustering* sebagai salah satu komponen dalam ranah yang lebih luas dalam ilmu *mechine learning*. Ilmu *mechine learning*, merupakan suatu proses kecerdasan buatan yang memungkinkan komputer

memperoleh pengetahuan dari data tanpa memerlukan pemrograman eksplisit (Géron 2019 dalam Jurczyk, T. 2021), sehingga suatu model *mechine learning*, setelah diinisiasi, mampu secara otomatis menemukan pola dalam data atau meramalkan data baru yang belum dikenal. Bidang ilmu *mechine learning* dapat dikelompokkan menjadi *supervised learning dan unsupervised learning* (Géron 2019 dalam Jurczyk, T. 2021).

a. *Supervised learning* menggunakan data yang diberi label untuk melatih algoritma guna membuat prediksi yang akurat untuk data baru. Misalnya, penyaringan spam yang memisahkan email sebagai "spam" atau "bukan spam." Evaluasi akurasi model melibatkan pengujian pada data berlabel dan penyesuaian parameter. Membangun model berkualitas melibatkan siklus pelatihan, pengujian, dan penyempurnaan parameter. Contoh klasifikasi umum adalah *k-nearest neighbors* (KNN) dan regresi logistik (Jurczyk, T. 2021).
b. *Unsupervised learning* digunakan pada data yang tak diberi label. Ini berguna untuk mendeteksi anomali, mengurangi dimensi data, dan pengelompokan data. Model tidak memerlukan data berlabel, melainkan mengenali pola dalam data. Misalnya, kita dapat menganalisis penulis atau ringkasan data dan mencari kelompok potensial tanpa referensi data berlabel. Penyetelan parameter seperti jumlah *cluster* dalam k-means dapat dievaluasi menggunakan metode seperti metode siku (*elbow method*) atau skor siluet (*Silhouette Score*) (Jurczyk, T. 2021).

2.2.6. *Elbow Method*

Karna, A., & Gibert, K. (2022) menyatakan bahwa *elbow method* adalah teknik yang digunakan untuk menentukan jumlah *cluster* optimal dalam analisis *cluster*. Metode ini memanfaatkan nilai inersia (*within-cluster sum of squares*) yang dihasilkan oleh algoritma klastering pada berbagai jumlah *cluster*. Tujuan dari *elbow method* adalah untuk mengidentifikasi titik di mana penambahan *cluster* tidak lagi memberikan pengurangan inersia yang signifikan, sehingga membentuk bentuk seperti "siku" pada grafik inersia. Langkah-langkah umum dari *elbow method* adalah sebagai berikut:

a. Jalankan algoritma klastering dengan berbagai nilai *cluster*.
b. Hitung inersia (*within-cluster sum of squares*) untuk setiap jumlah *cluster*.
c. Plot hasil inersia terhadap jumlah *cluster*.
d. Identifikasi titik di mana grafik menunjukkan titik "*elbow*" atau "siku".

Elbow method melibatkan konsep inersia dalam mengukur sejauh mana data pada sebuah *cluster* cenderung berdekatan satu sama lain. Semakin kecil nilai inersia, semakin baik. Namun, nilai inersia akan terus berkurang seiring penambahan jumlah *cluster*. *Elbow method* membantu menemukan jumlah *cluster* di mana peningkatan kualitas *cluster* tidak lagi sebanding dengan peningkatan jumlah *cluster*.

Konsep inersia (dalam konteks k-means) dijabarkan pada persamaan 2.3:

$$\text{Inersia} = \sum_{i=1}^{K} \sum_{j=1}^{n_k} \left\| x_{ij} - c_k \right\|^2 \qquad \text{.....(2.3)}$$

Keterangan persamaan 2.3:

K : jumlah *cluster*.

n_k : jumlah sampel dalam *cluster* ke-K.

x_{ij} : sampel ke- j dalam *cluster* ke-K.

c_k : pusat *cluster* ke-K.

2.2.7. *Davies-Bouldin Index*

Davies-Bouldin Index (DBI) adalah sebuah metrik evaluasi keberagaman antar-kluster dalam analisis *cluster*. Tujuan utamanya adalah untuk mengukur seberapa baik *cluster-cluster* yang dihasilkan oleh algoritma klastering tertentu. Semakin rendah nilai DBI, semakin baik performa klasteringnya (Glüge, S., dkk. 2020).

Rumus *davies-bouldin index* (DBI) terdapat pada persamaan 2.4 (Ros, F., Riad, R., & Guillaume, S. 2023):

$$\text{DB(k)} = \frac{1}{k}\sum_{i=1}^{k} \max_{\text{i} \neq \text{j}} \left\{\frac{Si+Sj}{d(\overline{x_i},\overline{x_j})}\right\} \qquad \text{.....(2.4)}$$

Keterangan persamaan 2.4:

k : jumlah *cluster*.

S_i : dispersi dalam *cluster* $_i$

$d(x_i,x_j)$:jarak antara pusat *cluster* $_i$ dan $_j$

DBI bekerja dengan cara membandingkan setiap *cluster* dengan *cluster* lainnya dan menghitung rasio antara jarak antar *cluster* dengan keberagaman dalam masing-masing *cluster*. Semakin kecil nilai DBI, semakin baik klasteringnya, karena itu menunjukkan bahwa *cluster*-klasternya lebih terpisah dan lebih homogen.

Menurut Vergani, A. A., & Binaghi, E. (2018) bahwa Indeks Davies-Bouldin memiliki keunggulan karena tidak memerlukan pengetahuan tentang label

sebenarnya (*ground truth*) dari data, sehingga cocok untuk *clustering* tanpa supervisi. DBI juga mempertimbangkan jarak antara centroid *cluster*. Jika *cluster* tersebar luas dan centroidnya jauh dari satu sama lain, nilai DBI akan tinggi. Sebaliknya, jika *cluster* kompak dan centroidnya berdekatan, nilai DBI akan rendah. Semakin mendekati nol nilai *Davies Bouldin Index (DBI),* semakin baik kualitas *cluster* yang dihasilkan dari proses pengelompokan menggunakan algoritma *clustering* (Drl, I. R., Chrisnanto, Y. H., & Umbara, F. R. 2022).

2.2.8. Shilhouette Score

Silhouette score adalah metode evaluasi yang digunakan untuk mengukur seberapa baik pemisahan *cluster* dalam data. Metrik ini memberikan nilai antara -1 hingga 1, di mana nilai yang lebih tinggi menunjukkan bahwa objek dalam sebuah *cluster* lebih dekat dengan objek dalam *cluster* yang sama daripada dengan objek dalam *cluster* lainnya (Athey, T. L., dkk. 2019).

Rumus untuk *Silhouette Score* adalah pada persamaan 2.5:

$$S = \frac{1}{N}\sum_{i=1}^{k} \frac{b_i - a_i}{max(a_i, b_i)} \qquad \text{.....(2.5)}$$

Keterangan persamaan 2.5:

N : jumlah total sampel.

a_i : jarak rata-rata antara sampel $_i$ dan semua sampel dalam *cluster* yang sama.

b_i : jarak rata-rata antara sampel $_i$ dan semua sampel dalam *cluster* terdekat yang berbeda dengan *cluster* tempat sampel $_i$ berada.

Dimana nilai S = [-1, +1], : - 1 = buruk, 0 = *indifferent,* 1 = bagus Ros, F., Riad, R., & Guillaume, S. 2023)

Untuk setiap objek, *silhouette score* menghitung rata-rata jarak antara objek tersebut dengan objek lain dalam *cluster* yang sama (a_i) dan rata-rata jarak antara objek tersebut dengan objek dalam *cluster* lain terdekat (b_i). *Silhouette score* kemudian dihitung sebagai $(b_i - a_i) / \max(a_i, b_i)$.

Silhouette score menggambarkan seberapa baik objek berada di dalam *cluster* yang sesuai dengannya dibandingkan dengan *cluster* lainnya. Nilai *Silhouette Score* yang mendekati 1 menunjukkan bahwa objek tersebut ter*cluster* dengan baik, dengan jarak yang kecil ke objek lain dalam *cluster* yang sama dan jarak yang besar ke objek dalam *cluster* lain. Sebaliknya, nilai negatif menunjukkan bahwa objek tersebut mungkin ditempatkan di *cluster* yang salah. Secara keseluruhan, nilai *Silhouette Score* yang tinggi menunjukkan *clustering* yang baik, sementara nilai negatif menunjukkan *clustering* yang buruk.

2.2.9. Python

Python adalah sebuah bahasa pemrograman yang diciptakan oleh Guido van Rossum pada tahun 1991. Python dikenal dengan filosofi desain yang menekankan keterbacaan kode, dan bahasa ini digunakan untuk pemrograman dalam berbagai skala. Python telah menjadi bahasa pemrograman yang telah diadopsi sebagai bahasa pengajaran oleh banyak departemen ilmu komputer terkemuka di Amerika Serikat. Selain itu, bahasa Python digunakan dalam berbagai bidang, termasuk dalam kecerdasan buatan, *deep learning*, dan pengembangan berbagai alat dan kerangka kerja yang efisien, seperti PyTorch, TensorFlow, dan Scikit-learn. Python juga digunakan dalam ujian sekunder komputer di China sejak tahun 2018. Dengan demikian, Python adalah bahasa pemrograman serbaguna yang sangat

penting dan populer di dunia komputasi saat ini (Zhang, X., et.al 2019).

a. Scikit-learn

Scikit-learn adalah sebuah modul Python yang menggabungkan berbagai algoritma *mechine learning* terkini untuk menangani masalah berukuran menengah dalam supervised learning maupun unsupervised learning. Poin pentingnya adalah kesederhanaan penggunaan, kinerja, penyediaan dokumentasi yang baik, dan konsistensi antarmuka pemrograman aplikasi (API) (Elshawi, R., & Sakr, S. 2020). Untuk menginstal versi terbaru (dengan pip): `pip install --upgrade scikit-learn`

b. Pandas

Dikutip dari dokumentasi pandas.pydata.org, pandas merupakan pustaka *open source* dengan lisensi BSD yang menyajikan struktur data dan perangkat analisis data yang efisien dan mudah digunakan untuk bahasa pemrograman Python. Pandas dapat diinstal melalui pip dari PyPI: `pip install pandas.` Ketika sedang berurusan dengan data yang berbentuk tabel, seperti data yang biasanya disimpan dalam spreadsheet atau basis data, pandas merupakan alat yang sangat sesuai. Pandas akan mempermudah eksplorasi, proses pembersihan, dan manipulasi data. Dalam konteks pandas, istilah yang digunakan untuk tabel data adalah "DataFrame."

In [1]: import pandas as pd

Untuk menggunakan paket pandas, impor paketnya. Dalam dokumentasi pandas, digunakan alias "pd" sebagai singkatan untuk pandas. Sehingga, penggunaan "pd" untuk memuat pandas dianggap sebagai praktik standar.

c. Matplotlib

Matplotlib adalah sebuah pustaka Python untuk menghasilkan visualisasi 2D yang dapat dicetak dengan kualitas publikasi dalam berbagai format, serta digunakan dalam beragam lingkungan, termasuk skrip Python, shell Python, IPython, notebook Jupyter, server aplikasi *web*, dan empat toolkit antarmuka pengguna grafis (Hunter J., et.al 2020 dalam dokumentasi matplotlib.org). Matplotlib dapat diunduh dan diinstal dari PyPI: python -mpip install matplotlib.