JITET (Jurnal Informatika dan Teknik Elektro Terapan)
Vol. 13 No. 1, pISSN: 2303-0577 eISSN: 2830-7062 http://dx.doi.org/10.23960/jitet.v13i1.5581

KLASTERISASI TRACER STUDY ALUMNI UNIVERSITAS XYZ MENGGUNAKAN ALGORITMA K-MEANS

Fabiyan Atha Fernaldy[1*], Amalia Anjani Arifiyanti[2], Dhian Satria Yudha Kartika[3]

[1,2,3]Universitas Pembangunan Nasional "Veteran" Jawa Timur; Jl. Rungkut Madya No.1, Kec. Gn. Anyar, Surabaya, Jawa Timur 60294; 031-870-6369

Received: 29 November 2024
Accepted: 14 Januari 2025
Published: 20 Januari 2025

Keywords:
K-Means;
Clustering;
Alumni.

Corespondent Email:
fabiyanaf09@gmail.com

Abstrak. Penelitian ini bertujuan untuk menganalisis dan mengelompokkan data alumni berdasarkan Indeks Prestasi Kumulatif (IPK) dan masa tunggu untuk mendapatkan pekerjaan menggunakan algoritma K-Means. Metode Elbow dan Silhouette Score diterapkan untuk menentukan jumlah cluster yang optimal. Hasil evaluasi menunjukkan bahwa untuk dataset yang dianalisis, jumlah cluster optimal untuk dataset pertama adalah tiga, sedangkan untuk dataset kedua adalah dua, dengan nilai Silhouette Score tertinggi masingmasing 0.497656 dan 0.502767. Deskripsi hasil clustering mengungkapkan perbedaan karakteristik antara cluster, di mana cluster dengan rata-rata IPK tertinggi memiliki masa tunggu terendah untuk mendapatkan pekerjaan. Temuan ini memberikan wawasan berharga bagi pengembangan kurikulum dan program bimbingan karir, serta meningkatkan pemahaman tentang pola karir alumni. Penelitian ini diharapkan dapat menjadi referensi untuk studi lebih lanjut dalam bidang analisis data dan pengembangan pendidikan.

Abstract. *This study aims to analyze and cluster alumni data based on the Grade Point Average (GPA) and the waiting time to obtain employment using the K-Means algorithm. The Elbow method and Silhouette Score are applied to determine the optimal number of clusters. Evaluation results indicate that for the analyzed datasets, the optimal number of clusters for the first dataset is three, while for the second dataset it is two, with the highest Silhouette Score values of 0.497656 and 0.502767, respectively. The description of the clustering results reveals differences in characteristics among clusters, where the cluster with the highest average GPA has the shortest waiting time to obtain employment. These findings provide valuable insights for curriculum development and career guidance programs, as well as enhancing the understanding of alumni career patterns. This research is expected to serve as a reference for further studies in the field of data analysis and educational development.*

1. PENDAHULUAN

Pendidikan tinggi memiliki peran penting dalam menyiapkan individu untuk menjadi anggota tenaga kerja yang kompeten dan mampu berkontribusi pada kemajuan masyarakat dan ekonomi [1]. Universitas Pembangunan Nasional "Veteran" Jawa Timur berkomitmen untuk mencetak lulusan berkualitas. Keberhasilan suatu program studi biasanya dilihat dari seberapa baik lulusannya bisa meraih kesuksesan dalam karir setelah menyelesaikan pendidikan.

Melalui tracer study, institusi pendidikan dapat memantau kondisi alumni setelah lulus. Data yang diperoleh dapat digunakan untuk merumuskan kebijakan dan langkah-langkah

yang mendukung alumni sekaligus membantu pengembangan institusi [2]. Universitas telah mengelola sistem Tracer Study sejak tahun 2021, dengan cakupan yang terus berkembang setiap tahunnya.

Para orang tua mahasiswa umumnya berharap lulusan perguruan tinggi memiliki Indeks Prestasi Kumulatif (IPK) yang baik, waktu tunggu yang singkat untuk mendapatkan pekerjaan, serta pekerjaan yang relevan dengan program studi yang diambil [3]. Oleh karena itu, evaluasi menyeluruh terhadap perkembangan karir alumni menjadi penting sebagai upaya meningkatkan kualitas pendidikan dan memberikan masukan berharga bagi universitas. Namun hingga saat ini, Universitas ini belum memiliki dokumentasi data terkait karakteristik karir alumni.

Sebuah penelitian membahas penggunaan algoritma K-Means untuk klasterisasi data hasil tracer study lulusan perguruan tinggi terkait karir dan pekerjaan. Penelitian ini menunjukkan bahwa analisis data tracer study berhasil mengelompokkan beberapa klaster lulusan, dengan skor evaluasi Davies-Bouldin Index (DBI) mencapai 0,287 pada percobaan pertama dan 0,291 pada percobaan kedua [3].

Penelitian ini bertujuan untuk melakukan klasterisasi data alumni melalui tracer study, yang merupakan alat penting untuk melacak perkembangan karir lulusan. Tracer study memungkinkan institusi untuk mengevaluasi sejauh mana alumni berhasil dalam karir mereka, termasuk waktu tunggu dalam mendapatkan pekerjaan dan hubungan antara bidang studi dengan pekerjaan yang diambil.

Dalam penelitian ini, hasil klasterisasi dibatasi pada dua aspek utama: pertama, klasterisasi berdasarkan Indeks Prestasi Kumulatif (IPK) dan waktu tunggu alumni untuk mendapatkan pekerjaan, yang memberikan wawasan tentang seberapa cepat alumni memperoleh pekerjaan sesuai dengan IPK mereka. Kedua, klasterisasi berdasarkan hubungan antara pekerjaan yang diambil dengan program studi yang ditempuh, yang membantu pihak Unit Pengembangan Karir dan Kewirausahaan (UPA-PKK) dalam memahami relevansi pendidikan yang diberikan dengan kebutuhan di dunia kerja.

Algoritma klasterisasi yang digunakan adalah K-Means, karena memiliki keunggulan dalam hal akurasi terhadap dimensi objek. Algoritma ini dianggap lebih terukur dan efisien untuk memproses data dengan jumlah objek yang besar. Selain itu, K-Means juga tidak sensitif terhadap urutan data yang dimasukkan [4]. Dengan hasil klasterisasi ini, diharapkan UPA-PKK dapat lebih efektif dalam merencanakan program pengembangan karir bagi alumni, serta meningkatkan mutu pendidikan yang diselenggarakan.

2. TINJAUAN PUSTAKA

2.1. Tracer Study

Tracer study atau survei alumni adalah penelitian yang bertujuan untuk melacak perjalanan lulusan dari perguruan tinggi. Studi ini dilakukan untuk memahami berbagai aspek hasil pendidikan, seperti transisi dari dunia pendidikan tinggi ke dunia kerja, penilaian lulusan terhadap keterampilan yang diperoleh selama studi, evaluasi proses pembelajaran, serta dampak pendidikan tinggi terhadap penguasaan keterampilan. Selain itu, tracer study juga mencakup pengumpulan informasi tambahan terkait lulusan [5].

2.2. Data Mining

Data mining adalah proses yang hampir otomatis yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan machine learning untuk mengambil dan menemukan informasi yang berguna serta mungkin tersembunyi dalam kumpulan data yang besar [6]. Proses ini dimulai dengan pencarian dan analisis data dalam jumlah besar untuk menemukan pola atau informasi menarik menggunakan metode tertentu. Pemilihan teknik atau algoritma yang digunakan disesuaikan dengan tujuan dan keseluruhan proses penemuan pengetahuan dalam basis data. Tahapan ini menjadi inti dari proses penemuan pengetahuan, dilakukan untuk menganalisis data yang sebelumnya telah dibersihkan [7].

2.3. Clustering

Salah satu metode pengelompokan dalam data mining adalah clustering. Clustering bertujuan untuk mengelompokkan data atau objek ke dalam beberapa kelompok, di mana setiap kelompok terdiri dari data yang serupa, tetapi berbeda dari data di kelompok lain. Metode ini digunakan untuk mengelompokkan

data berdasarkan kesamaan atau perbedaan tertentu dengan data di klaster lainnya [8].

2.4. Algoritma K-Means

K-Means adalah algoritma unsupervised learning yang digunakan untuk mengelompokkan data berdasarkan kesamaan atau kemiripan [9]. Algoritma ini termasuk dalam metode non-hierarki yang berfungsi untuk mengelompokkan data ke dalam beberapa cluster, di mana setiap cluster memiliki karakteristik yang serupa dan berbeda dari cluster lainnya. Dalam konteks unsupervised learning, K-Means Clustering digunakan untuk menganalisis data dan membentuk partisi berdasarkan karakteristik yang mirip di antara data yang ada [10].

K-Means juga termasuk dalam algoritma partitional, karena algoritma ini bekerja dengan membagi data menjadi sejumlah kelompok yang telah ditentukan sebelumnya. Prosesnya dimulai dengan mendefinisikan nilai centroid awal untuk setiap kelompok, kemudian data dikelompokkan berdasarkan kedekatannya dengan centroid tersebut. Setiap iterasi mengubah posisi centroid sampai konvergensi tercapai, yaitu ketika posisi centroid tidak lagi berubah secara signifikan [11].

2.5. Python

Python adalah bahasa pemrograman tingkat tinggi yang terkenal karena kemudahan penggunaannya. Karena alasan ini, banyak programmer memilih Python untuk mengembangkan program mereka [12]. Dengan sintaks yang sederhana, Python juga dianggap mudah dipelajari. Selain itu, Python dilengkapi dengan berbagai pustaka yang lengkap dan didukung oleh komunitas yang kuat karena sifatnya yang open source.

Saat menulis kode Python, pengguna dapat menggunakan berbagai Integrated Development Environment (IDE) seperti VS Code, Sublime Text, PyCharm, atau bahkan IDE online seperti Jupyter Notebook dan Google Colab [13].

2.6. iVAT

Improved Visual Assessment for Tendency (iVAT) adalah versi yang lebih baik dari algoritma VAT (Visual Assessment of Tendency). iVAT merupakan teknik visualisasi data yang digunakan untuk menganalisis cluster dengan menggambarkan matriks jarak antara objek dalam cluster, memudahkan dalam mengidentifikasi pola dan struktur dalam data [14].

iVAT menghasilkan kotak atau persegi hitam yang lebih tepat dan jelas pada peta cetaknya. Dalam situasi jarang, gambar persegi hitam yang mewakili jumlah cluster tidak akan terletak di diagonal kiri dan mungkin tidak berbentuk persegi [15]. Algoritma iVAT menawarkan visualisasi yang lebih berguna dan lebih efektif dibandingkan dengan VAT dalam mengidentifikasi kecenderungan cluster, terutama dalam kasus yang lebih kompleks di mana VAT mungkin tidak memberikan hasil yang memadai [16].

2.7. Elbow Method

Metode elbow adalah teknik yang digunakan untuk menentukan jumlah k optimal dalam pembuatan klaster [17]. Metode ini bekerja dengan memvisualisasikan berbagai jumlah klaster dan menghitung nilai klaster yang akan digunakan sebagai model data untuk menentukan klaster terbaik. Selain itu, persentase perhitungan yang dihasilkan merupakan perbandingan antara jumlah klaster yang ditambahkan, membantu untuk mengidentifikasi titik di mana penambahan klaster tidak lagi memberikan perbaikan signifikan pada model.

Perbedaan persentase dari setiap nilai klaster dapat digambarkan dalam grafik sebagai sumber informasi. Jika nilai klaster pertama dan kedua membentuk sudut atau menunjukkan penurunan terbesar pada grafik, maka nilai klaster tersebut dianggap sebagai yang terbaik. Untuk perbandingan, SSE (Jumlah Kesalahan Kuadrat) dihitung untuk setiap nilai klaster. Semakin besar nilai K, semakin kecil nilai SSE yang dihasilkan [18].

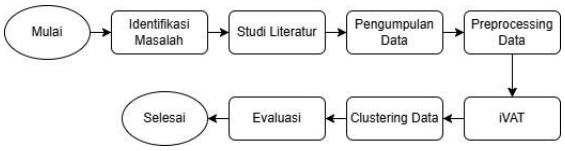
2.8. Silhouette Coefficient

Silhouette Coefficient merupakan metrik yang digunakan untuk mengevaluasi seberapa baik teknik pengelompokan bekerja, dengan mengukur sejauh mana setiap objek cocok dengan klasternya dibandingkan dengan klister lain, dengan skala nilai antara -1 dan 1 [19]. Nilai tinggi menunjukkan kesesuaian yang baik antara objek dan klasternya, namun kurang sesuai dengan klaster tetangga. Sebaliknya, nilai rendah atau negatif menunjukkan bahwa mungkin terdapat terlalu banyak atau terlalu

sedikit klaster dalam konfigurasi pengelompokan [20].

3. METODE PENELITIAN

Dalam bab ini diuraikan alur atau tahapan-tahapan yang akan digunakan dalam penelitian, sehingga penelitian dapat dilakukan dengan baik dan terstruktur.



Gambar 1. Diagram Alir Penelitian

3.1. Identifikasi Masalah

Dalam penelitian ini, dilakukan identifikasi masalah mengenai tujuan dari masalah yang ingin diselesaikan. Proses identifikasi masalah dilakukan dengan wawancara dengan pihak Unit Pengembangan Karir dan Kewirausahaan (UPA-PKK) Universitas.

3.2. Studi Literatur

Selanjutnya, dilakukan studi literatur yang bertujuan untuk mengumpulkan dan menganalisis berbagai literatur yang relevan dengan penelitian ini. Sumber-sumber untuk studi literatur tersebut diperoleh dari beberapa jurnal dan artikel yang berkaitan dengan data mining, tracer study, dan K-Means clustering.

3.3. Pengumpulan Data

Data yang akan digunakan dalam penelitian ini diperoleh dari koordinator tracer study UPA-PKK UPN "Veteran" Jawa Timur. Data lulusan yang digunakan mencakup lulusan mahasiswa dari tahun 2021 hingga 2022, dengan total sebanyak 5.313 data. Selain itu, data kuesioner tracer study disajikan dalam bentuk file excel workbook (xlsx). Dari tujuan penelitian ini, atribut yang digunakan klasterisasi adalah IPK, f502 (waktu tunggu alumni untuk mendapatkan perkajaan), dan f14 (hubungan pekerjaan dengan program studi).

3.4. Preprocessing Data

Pada tahap ini juga dilakukan cleansing data dengan membersihkan data redundant (duplikasi) dan data yang bernilai null sehingga data siap untuk digunakan dalam tahap selanjutnya. Tahapan awal yaitu pemilihan fitur.

```
df_1_selected = df[['Last Modified','kdptesmh', 'nimhsmsmh', 'f8', 'f1101','ipk', 'f502', 'tahun_lulus','f5a1', 'f5a2', 'f505']]
df_2_selected = df[['Last Modified','kdptesmh', 'nimhsmsmh', 'f8', 'f1101', 'f502', 'f14', 'tahun_lulus','f5a1', 'f5a2', 'f505']]

print("DataFrame Pertama:")
print(df_1_selected.head())

print("\nDataFrame Kedua:")
print(df_2_selected.head())
```

Gambar 2. Source Code Tahap Pemilihan Fitur

Gambar 2 menunjukkan proses pemilihan data dari dua DataFrame yang berbeda. DataFrame pertama bernama df_1_selected dan DataFrame kedua bernama df_2_selected. DataFrame df_1_selected digunakan untuk klasterisasi alumni berdasarkan IPK dan masa tunggu (f502), sedangkan df_2_selected digunakan untuk mengelompokkan alumni berdasarkan hubungan antara pekerjaan dan program studi (f14) dengan masa tunggu (f502).

```
DataFrame Pertama:
  Last Modified  kdptesmh  nimhsmsmh   f8  f1101   ipk  f502  tahun_lulus  \
0    17/11/2022     22201  1453010001  5.0    0.0  3.42   0.0         2021
1    17/11/2022     22201  1453010002  1.0    5.0  3.42   3.0         2021
2    13/05/2022     22201  1453010002  NaN    NaN  3.42   NaN         2021
3    17/11/2022     22201  1453010021  1.0    3.0  3.42   4.0         2021
4    13/05/2022     22201  1453010021  NaN    NaN  3.42   NaN         2021

     f5a1     f5a2      f505
0  50000.0  51400.0       0.0
1  50000.0  50200.0  4200000.0
2      NaN      NaN       NaN
3  50000.0  56000.0  4500000.0
4      NaN      NaN       NaN

DataFrame Kedua:
  Last Modified  kdptesmh  nimhsmsmh   f8  f1101  f502  f14  tahun_lulus  \
0    17/11/2022     22201  1453010001  5.0    0.0   0.0  0.0         2021
1    17/11/2022     22201  1453010002  1.0    5.0   3.0  3.0         2021
2    13/05/2022     22201  1453010002  NaN    NaN   NaN  NaN         2021
3    17/11/2022     22201  1453010021  1.0    3.0   4.0  1.0         2021
4    13/05/2022     22201  1453010021  NaN    NaN   NaN  NaN         2021

     f5a1     f5a2      f505
0  50000.0  51400.0       0.0
1  50000.0  50200.0  4200000.0
2      NaN      NaN       NaN
3  50000.0  56000.0  4500000.0
4      NaN      NaN       NaN
```

Gambar 3. Hasil Source Code Tahap Pemilihan Fitur

Tahapan berikutnya adalah Data Cleansing. Pertama, data difilter dengan menggunakan metode .isin([1, 3]) untuk mempertahankan baris yang memiliki nilai 'f8' sama dengan 1 atau 3. Nilai 1 dalam 'f8' menunjukkan alumni yang bekerja full time/part time, sedangkan nilai 3 menunjukkan alumni yang bekerja wiraswasta.

Dilakukan pembersihan lanjutan pada dua DataFrame df_1 dan df_2.

```
df_1 = df_1.dropna(subset=['f502'])

df_1['Last Modified'] = pd.to_datetime(df_1['Last Modified'], format='%d/%m/%Y')

df_1 = df_1.sort_values(['nimhsmsmh', 'Last Modified'], ascending=[True, False])

df_1_cleaned = df_1.drop_duplicates(subset=['nimhsmsmh'], keep='first')
df_1_cleaned.info()
df_2 = df_2.dropna(subset=['f502'])

df_2['Last Modified'] = pd.to_datetime(df_2['Last Modified'], format='%d/%m/%Y')

df_2 = df_2.sort_values(['nimhsmsmh', 'Last Modified'], ascending=[True, False])

df_2_cleaned = df_2.drop_duplicates(subset=['nimhsmsmh'], keep='first')
df_2_cleaned.info()
```

Gambar 4. Source Code Tahap Penanganan Nilai Yang Dan Duplikat Data Frame

Tahapan berikutnya adalah melakukan analisis korelasi antara atribut IPK, f502 (waktu tunggu alumni untuk mendapatkan pekerjaan), dan f14 (hubungan pekerjaan dengan program studi). Koefisien korelasi merupakan angka atau indeks yang mengukur seberapa kuat hubungan antara dua variabel.

Selanjutnya, dilakukan tahap normalisasi data. Tahapan ini penting karena atribut data yang digunakan memiliki perbedaan dalam dimensi. Normalisasi data adalah proses penskalaan nilai atribut dari suatu data sehingga data dapat terletak pada rentang skala tertentu [21].

```
df1_normalisasi = pd.DataFrame(df_1_cleaned[['ipk', 'f502']])
minmax_scaler = preprocessing.MinMaxScaler()
X_minmax = minmax_scaler.fit_transform(df1_normalisasi)
df1_normalized = pd.DataFrame(X_minmax, columns=df1_normalisasi.columns)

df1_normalized
df2_normalisasi = pd.DataFrame(df_2_cleaned[['f502', 'f14']])
minmax_scaler = preprocessing.MinMaxScaler()
X_minmax = minmax_scaler.fit_transform(df2_normalisasi)
df2_normalized = pd.DataFrame(X_minmax, columns=df2_normalisasi.columns)

df2_normalized
```

Gambar 5. Source Code Tahap Normalisasi Data Frame 1 dan 2

Metode normalisasi yang digunakan dalam penelitian ini adalah Min-Max normalization. Metode normalisasi Min-Max adalah teknik yang mengubah rentang nilai data sehingga berada dalam kisaran antara 0 dan 1. Persamaan untuk menghitung MinMax Normalization dapat dilihat pada persamaan berikut [22].

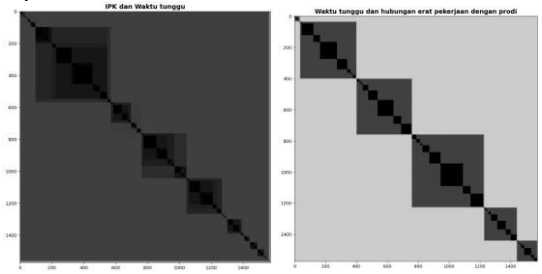
$$x' = \frac{x_i - \min(x)}{\max(x) - \min(x)} \quad (1)$$

Keterangan:

x_i= nilai tertentu yang akan dinormalisasi
x'= nilai hasil normalisasi
min (x) = nilai minimal dari sebuah atribut
max (x) = nilai maksimal dari sebuah atribut

3.5 Visualisasi Tendency Cluster (iVAT)

Tahap ini bertujuan untuk mengidentifikasi kemungkinan adanya cluster dengan menggunakan iVAT (Improved Visual Analysis for Cluster Tendency). Dalam gambar matriks iVAT, cluster akan terlihat sebagai blok diagonal gelap yang lebih jelas dan terpisah.



Gambar 6. Hasil Source Code Tahap Visualisasi Tendency Cluster Data Frame 1 dan Data Frame 2

3.6 Clustering Data

Pada tahap clustering data, data akan dilakukan pemodelan cluster dengan menggunakan bahasa pemrograman python. Langkah yang dilakukan untuk membentuk clustering pada metode K-Means sebagai berikut [23].

1. Untuk menentukan jumlah cluster (k) yang optimal dari dataset, kita dapat menggunakan metode siku (elbow method). Dalam metode ini, kita mencari nilai cluster terbaik berdasarkan penurunan signifikan pada Sum of Square Error (SSE), yang akan membentuk pola siku pada grafik. SSE dihitung menggunakan rumus tertentu untuk mengukur seberapa baik data dikelompokkan dalam cluster [18].
$$SSE = \sum_{i=1}^{n}(d)^2 \quad (2)$$
2. Menentukan k sebagai Centroid, dilakukan secara acak (random).
3. Hitung jarak data dengan centroid menggunakan rumus jarak menggunakan rumus Euclidean.
$$D_e = \sqrt{(x_i - s_i)^2 + (y_i - t_i)^2} \quad (3)$$
4. Menghitung pusat cluster baru dilakukan dengan menentukan rata-rata dari semua anggota dalam setiap cluster. Centroid baru ini mencerminkan posisi tengah dari setiap kelompok data.
5. Menghitung ulang setiap objek dengan menggunakan centroid baru hingga anggota cluster tidak berubah lagi. Jika masih ada perubahan, ulangi langkah 3 dan 4 sampai tidak ada perubahan pada anggota cluster.

3.7 Evaluasi

Pada tahap ini, mengevaluasi keberhasilan pembentukan cluster menggunakan Koefisien Siluet. Koefisien ini dihitung dengan mempertimbangkan jarak rata-rata dalam cluster (a) dan jarak rata-rata ke cluster terdekat (b) untuk setiap titik data, menggunakan rumus (b - a)/max(a, b) [24].

- Skor siluet mendekati +1 menunjukkan bahwa titik data berada di cluster yang tepat.
- Skor siluet mendekati 0 menunjukkan bahwa titik data mungkin cocok dengan cluster lain.

- Skor siluet mendekati -1 menunjukkan bahwa titik data berada di cluster yang salah.

4. HASIL DAN PEMBAHASAN

4.1 Clustering Data

Tahap yang pertama adalah menentukan nilai K yang optimal.

```
def elbow_method(data, max_k=10):
    sse = []
    for k in range(1, max_k + 1):
        kmeans = KMeans(n_clusters=k, random_state=42)
        kmeans.fit(data)
        sse.append(kmeans.inertia_)
```

Gambar 7. Source Code Tahap Perhitungan SSE untuk K-means Clustering

Pada gambar 7, terdapat fungsi bernama elbow_method yang digunakan untuk menentukan nilai k optimal. Fungsi ini menerima dua parameter: ‘data’ (dataset yang akan dianalisis) dan ‘max_k’ (jumlah maksimum klaster yang akan diuji).

Pertama, sebuah list kosong bernama sse (Sum of Squared Errors) diinisialisasi untuk menyimpan nilai error dari setiap jumlah klaster. Selanjutnya, dilakukan iterasi dari 1 hingga max_k, di mana pada setiap iterasi, model KMeans dibuat sesuai jumlah klaster k. Model ini kemudian di-fit ke ‘data’, dan nilai inertia (yang merepresentasikan SSE) dari model tersebut ditambahkan ke dalam list SSE.

```
# Membuat tabel SSE
sse_table = pd.DataFrame({
    'Jumlah Cluster': range(1, max_k + 1),
    'SSE': sse
})
print("Tabel SSE:")
print(sse_table.to_string(index=False))
```

Gambar 8. Source Code Tahap Membuat Tabel SSE

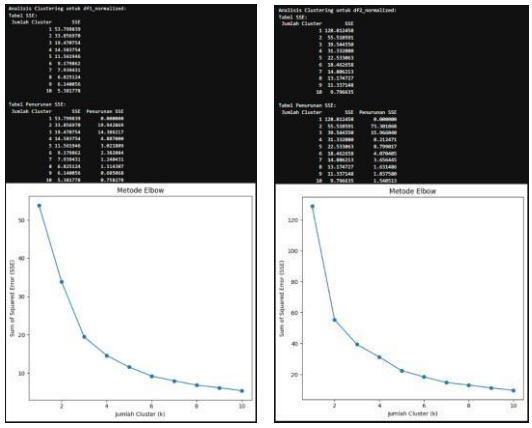
Pada gambar 8, sebuah DataFrame pandas dibuat dengan nama **sse_table**. DataFrame ini memiliki dua kolom: 'Jumlah Cluster', yang berisi rentang angka dari 1 hingga **max_k** (jumlah maksimum klaster yang diuji), dan 'SSE', yang berisi nilai-nilai SSE yang telah dihitung sebelumnya dan disimpan dalam list **sse**. Setelah DataFrame dibuat, kode mencetak label "Tabel SSE:" diikuti dengan menampilkan seluruh isi tabel tanpa nomor indeks baris, sehingga pengguna dapat dengan mudah melihat hubungan antara jumlah cluster dan nilai SSE yang dihasilkan.

```
# Menghitung penurunan SSE
sse_decrease = [0] + [sse[i-1] - sse[i] for i in range(1, len(sse))]
sse_table['Penurunan SSE'] = sse_decrease
print("\nTabel Penurunan SSE:")
print(sse_table.to_string(index=False))
```

Gambar 9. Source Code Tahap Menghitung Penurunan SSE

```
plt.figure(figsize=(12, 6))
plt.plot(range(1, max_k + 1), sse, marker='o')
plt.xlabel('Jumlah Cluster (k)')
plt.ylabel('Sum of Squared Error (SSE)')
plt.title('Metode Elbow')
plt.show()
```

Gambar 10. Source Code Tahap Visualisasi Grafik Metode Elbow



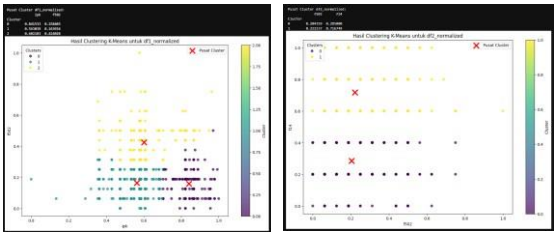
Gambar 11. Source Code Tahap Penerapan Fungsi elbow_method

Tahap berikutnya yaitu assign hasil klaster.

```
# Visualisasi hasil clustering
fig, ax = plt.subplots(figsize=(12, 8))
scatter = ax.scatter(df1_normalized.iloc[:, 0], df1_normalized.iloc[:, 1], c=labels_df1, cmap='viridis', alpha=0.7)
ax.scatter(cluster_centers1[:, 0], cluster_centers1[:, 1], c='red', marker='x', s=200, linewidths=3, label='Pusat Cluster')
ax.set_xlabel(df1_normalized.columns[0])
ax.set_ylabel(df1_normalized.columns[1])
ax.set_title('Hasil Clustering K-Means untuk df1_normalized')
legend1 = ax.legend(*scatter.legend_elements(), title="Clusters", loc="upper left")
ax.add_artist(legend1)
ax.legend(loc='upper right')
plt.colorbar(scatter, label='Cluster')
plt.show()
```

```
# Visualisasi hasil clustering
fig, ax = plt.subplots(figsize=(12, 8))
scatter2 = ax.scatter(df2_normalized.iloc[:, 0], df2_normalized.iloc[:, 1], c=labels_df2, cmap='viridis', alpha=0.7)
ax.scatter(cluster_centers2[:, 0], cluster_centers2[:, 1], c='red', marker='x', s=200, linewidths=3, label='Pusat Cluster')
ax.set_xlabel(df2_normalized.columns[0])
ax.set_ylabel(df2_normalized.columns[1])
ax.set_title('Hasil Clustering K-Means untuk df2_normalized')
legend2 = ax.legend(*scatter2.legend_elements(), title="Clusters", loc="upper left")
ax.add_artist(legend2)
ax.legend(loc='upper right')
plt.colorbar(scatter2, label='Cluster')
plt.show()
```

Gambar 12. Source Code Tahap Visualisasi Hasil Clustering



Gambar 13. Hasil Source Code Tahap Analisis Pusat Cluster dan Tahap Visualisasi Hasil Clustering

Gambar 14 menampilkan hasil clustering dari dataset df1_normalized dan df2_normalized.

```
Hasil Clustering df1_normalized:
        ipk    f502  Cluster
0  0.585185  0.1875        0
1  0.585185  0.2500        0
2  0.585185  0.2500        0
3  0.533333  0.1250        0
4  0.533333  0.1875        0
5  0.533333  0.3125        2
6  0.474074  0.1250        0
7  0.474074  0.2500        0
8  0.474074  0.1875        0
9  0.474074  0.5000        2

Jumlah data per cluster:
0    968
1    298
2    306
Name: Cluster, dtype: int64
```

```
Hasil Clustering df2_normalized:
     f502  f14  Cluster
0  0.1875  0.2        0
1  0.0625  0.2        0
2  0.2500  0.8        1
3  0.7500  0.6        1
4  0.0625  0.6        1
5  0.1875  1.0        1
6  0.1875  0.8        1
7  0.1875  0.6        1
8  0.1875  0.4        0
9  0.0625  0.4        0

Jumlah data per cluster:
0    760
1    812
Name: Cluster, dtype: int64
```

Gambar 14. Hasil Source Code Tahap Menampilkan Distribusi Data Tiap Cluster

4.2 Deskripsi Hasil Clustering

Berikut adalah deskripsi hasil pengelompokkan data alumi Universitas pada Tracer Study.

1. Masa tunggu alumni mendapatkan pekerjaan dengan IPK memiliki 3 cluster.
   - Cluster 0 memiliki rata-rata IPK tertinggi dari pada klaster lainnya, rata-rata di 3.77 dengan rentang IPK 3.58 - 3.98 serta memiliki Masa tunggu untuk mendapatkan perkerjaan rata-rata di 2.53 bulan (terendah di antara ketiga klaster) dengan rentang 0 - 8 bulan.
   - Cluster 1 memiliki rata-rata IPK terendah dari pada klaster lainnya, rata-rata di 3.39 dengan rentang IPK 2.63 - 3.58 serta memiliki Masa tunggu untuk mendapatkan perkerjaan rata-rata di 2.62 bulan (menengah di antara ketiga klaster) dengan rentang 0 - 5 bulan.
   - Cluster 2 memiliki rata-rata IPK menengah dari pada klaster lainnya, rata-rata di 3.44 dengan rentang IPK 3.11 - 3.93 serta memiliki Masa tunggu untuk mendapatkan perkerjaan rata-rata di 6.78 bulan (tertinggi di antara ketiga klaster) dengan rentang 5 - 16 bulan.
2. Masa tunggu alumni mendapatkan perkerjaan dengan IPK berdasarkan jenis program studi alumni memiliki 3 cluster.
   - Cluster 0, Alumni dalam cluster ini cenderung bekerja di sektor Perusahaan swasta sebanyak 153 alumni. Cluster ini memiliki persentase signifikan terbesar di Institusi/Organisasi Multilateral sebesar 41.8% dari total alumni yang bekerja di Institusi/Organisasi Multilateral.
   - Cluster 1, Alumni dalam cluster ini cenderung bekerja di sektor Perusahaan swasta sebanyak 399 alumni. Cluster ini memiliki persentase signifikan terbesar di BUMN/BUMD sebesar 74.1% dari total alumni yang bekerja di BUMN/BUMD
   - Cluster 2, Alumni dalam cluster ini cenderung bekerja di sektor Perusahaan swasta sebanyak 127 alumni. Cluster ini memiliki persentase signifikan terbesar di Organisasi non-profit/Lembaga Swadaya Masyarakat sebesar 42.1% dari total alumni yang bekerja di Organisasi non-profit/Lembaga Swadaya Masyarakat
3. Masa tunggu alumni mendapatkan perkerjaan dengan IPK berdasarkan jenis pekerjaan alumni memiliki 3 cluster.
   - Cluster 0, cenderung didominasi oleh program studi Akuntansi (36 alumni). Selain itu terdapat program studi yang memiliki persentase yang signifikan dibanding cluster lain, program studi Magister Agroteknologi sebesar 100.0% dari total alumni Magister Agroteknologi yang telah bekerja.
   - Cluster 1, cenderung didominasi oleh program studi Manajemen (132 alumni). Selain itu terdapat program studi yang memiliki persentase yang signifikan dibanding cluster lain, program studi Agribisnis sebesar 83.0% dari total alumni Agribisnis yang telah bekerja.
   - Cluster 2, cenderung didominasi oleh program studi Akuntansi (42 alumni). Selain itu terdapat program studi yang memiliki persentase yang signifikan dibanding cluster lain, program studi Teknik Kimia sebesar

26.6% dari total alumni Teknik Kimia yang telah bekerja.

4. Masa tunggu alumni mendapatkan pekerjaan dengan hubungan erat pekerjaan - prgoram studi memiliki 2 cluster.
   - Cluster 0 mencakup alumni yang memiliki hubungan Sangat Erat, dan Erat serta menunjukkan karakteristik rata-rata masa tunggu alumni mendapatkan perkerjaan sebesar 3.27 bulan dengan rentang 0-12 bulan.
   - Cluster 1 mencakup alumni yang memiliki hubungan Cukup Erat, Kurang Erat, dan Tidak Sama Sekali serta menunjukkan karakteristik ratarata masa tunggu alumni mendapatkan perkerjaan sebesar 3.55 bulan dengan rentang 0-16 bulan.
5. Masa tunggu alumni mendapatkan perkerjaan dengan hubungan erat pekerjaan - prgoram studi berdasarkan jenis pekerjaan alumni memiliki 2 cluster.
   - Cluster 0, alumni dalam cluster ini cenderung bekerja di sektor Perusahaan swasta sebanyak 341 alumni. Cluster ini memiliki persentase signifikan terbesar di Instansi pemerintah sebesar 55.9% dari total alumni yang bekerja di Instansi pemerintah.
   - Cluster 1, alumni dalam cluster ini cenderung bekerja di sektor Perusahaan swasta sebanyak 338 alumni. Cluster ini memiliki persentase signifikan terbesar di Organisasi non-profit/Lembaga Swadaya Masyarakat sebesar 63.2% dari total alumni yang bekerja di Organisasi non-profit/Lembaga Swadaya Masyarakat.
6. Masa tunggu alumni mendapatkan perkerjaan dengan hubungan erat pekerjaan - prgoram studi berdasarkan jenis program studi alumni memiliki 2 cluster.
   - Cluster 0, cenderung didominasi oleh program studi Akuntansi (119 alumni). Selain itu terdapat program studi yang memiliki persentase yang signifikan dibanding cluster lain, program studi Magister Akuntansi sebesar 81.8% dari total alumni Magister Akuntansi yang telah bekerja.
   - Cluster 1, cenderung didominasi oleh program studi Manajemen (82 alumni). Selain itu terdapat program studi yang memiliki persentase yang signifikan dibanding cluster lain, program studi Magister Ilmu Lingkungan sebesar 100.0% dari total alumni Magister Ilmu Lingkungan yang telah bekerja.

4.3 Evaluasi

```python
def evaluate_clustering(data, max_k=10):
    silhouette_scores = []
    for k in range(2, max_k + 1):
        kmeans = KMeans(n_clusters=k, random_state=42)
        labels = kmeans.fit_predict(data)
        score = silhouette_score(data, labels)
        silhouette_scores.append(score)

    # Menambahkan NaN untuk k=1
    silhouette_scores = [float('nan')] + silhouette_scores

    # Membuat tabel Silhouette Score
    silhouette_table = pd.DataFrame({
        'Cluster': range(1, max_k + 1),
        'Silhouette Score': silhouette_scores
    })

    print("Tabel Silhouette Score df1_normalized:")
    print(silhouette_table.to_string(index=False))

evaluate_clustering(df1_normalized)
```

```python
def evaluate_clustering(data, max_k=10):
    silhouette_scores = []
    for k in range(2, max_k + 1):
        kmeans = KMeans(n_clusters=k, random_state=42)
        labels = kmeans.fit_predict(data)
        score = silhouette_score(data, labels)
        silhouette_scores.append(score)

    # Menambahkan NaN untuk k=1
    silhouette_scores = [float('nan')] + silhouette_scores

    # Membuat tabel Silhouette Score
    silhouette_table = pd.DataFrame({
        'Cluster': range(1, max_k + 1),
        'Silhouette Score': silhouette_scores
    })

    print("Tabel Silhouette Score df2_normalized:")
    print(silhouette_table.to_string(index=False))

evaluate_clustering(df2_normalized)
```

Gambar 15. Source Code Tahap Evaluasi Cluster

Gambar 15 mendefinisikan fungsi **evaluate_clustering**, yang bertujuan untuk mengevaluasi kualitas clustering menggunakan metode Silhouette Score. Fungsi ini menerima dua parameter: dataset dan jumlah maksimum cluster. Fungsi ini kemudian melakukan iterasi dari 1 hingga jumlah maksimum cluster yang ditentukan. Pada setiap iterasi, algoritma K-means dijalankan, Silhouette Score dihitung, dan hasilnya disimpan.

```
Tabel Silhouette Score df1_normalized:
 Cluster  Silhouette Score
       1               NaN
       2          0.483205
       3          0.499046
       4          0.447196
       5          0.485410
       6          0.500263
       7          0.489857
       8          0.474558
       9          0.453080
      10          0.481338
```

```
Tabel Silhouette Score df2_normalized:
 Cluster  Silhouette Score
       1               NaN
       2          0.502767
       3          0.447670
       4          0.454013
       5          0.499143
       6          0.524789
       7          0.557708
       8          0.546588
       9          0.578386
      10          0.581966
```

Gambar 16. Hasil Source Code Tahap Evaluasi Cluster

Gambar 16 menampilkan tabel Silhouette Score untuk hasil clustering dengan jumlah cluster yang berbeda-beda. Untuk cluster dengan jumlah 1, nilainya adalah NaN (Not a Number) karena Silhouette Score tidak dapat dihitung untuk satu cluster. Dalam tabel Silhouette Score untuk df1_normalized, skor tertinggi dicapai pada 3 cluster dengan nilai 0.497656, yang menunjukkan bahwa jumlah cluster yang optimal untuk dataset ini adalah 3.

Sementara itu, untuk tabel Silhouette Score **df2_normalized**, hasil menunjukkan bahwa 2 cluster adalah pilihan yang paling optimal. Silhouette Score untuk (k=2) memiliki skor tertinggi dengan nilai 0.502767, yang lebih baik dibandingkan dengan jumlah cluster lainnya. Dengan demikian, analisis ini memberikan wawasan yang berguna untuk menentukan jumlah cluster yang paling sesuai untuk masing-masing dataset.

5. KESIMPULAN

Penelitian ini berhasil menerapkan algoritma K-means untuk menganalisis dan mengelompokkan data alumni. Dengan menggunakan metode Elbow dan Silhouette Score, dapat menentukan jumlah cluster yang optimal untuk masing-masing dataset.

Hasil evaluasi menunjukkan bahwa untuk dataset **df1_normalized**, jumlah cluster yang optimal adalah 3 dengan Silhouette Score tertinggi sebesar 0.497656. Sedangkan untuk dataset **df2_normalized**, jumlah cluster yang optimal adalah 2 dengan Silhouette Score tertinggi sebesar 0.502767. Ini menunjukkan bahwa pemilihan jumlah cluster yang tepat dapat meningkatkan kualitas hasil clustering.

Hasil clustering menunjukkan adanya perbedaan karakteristik antara setiap cluster. Misalnya, analisis berdasarkan IPK dan masa tunggu alumni untuk mendapatkan pekerjaan menunjukkan bahwa cluster dengan rata-rata IPK tertinggi memiliki masa tunggu terendah, sementara cluster dengan IPK terendah memiliki masa tunggu yang lebih panjang.

Temuan ini memberikan wawasan yang berguna bagi pihak universitas dalam memahami pola karir alumni dan dapat digunakan untuk pengembangan kurikulum yang lebih baik serta meningkatkan program bimbingan karir bagi mahasiswa. Penelitian selanjutnya dapat berfokus pada metode clustering lainnya juga dipertimbangkan untuk analisis lebih lanjut, serta eksplorasi variabel tambahan yang mungkin mempengaruhi hasil clustering, guna mendapatkan pemahaman yang lebih komprehensif tentang data alumni.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada pihak UPA-PKK Universitas yang telah menyediakan data dan fasilitas yang diperlukan untuk pelaksanaan penelitian ini.

DAFTAR PUSTAKA

[1] I. W. C. Sujana, “FUNGSI DAN TUJUAN PENDIDIKAN INDONESIA,” *Adi Widya: Jurnal Pendidikan Dasar*, vol. 4, no. 1, p. 29, Jul. 2019, doi: 10.25078/aw.v4i1.927.

[2] V. B. Siahaan and A. R. Kardian, “Penerapan Algoritma K-Means Untuk Analisis Tracer Alumni Universitas Gunadarma Jurusan Sistem Informasi dan Sistem Komputer Angkatan 2013,” *Jurnal Ilmiah KOMPUTASI*, vol. 18, no. 3, pp. 215–228, Sep. 2019.

[3] Joko Sutrisno, Arief Wibowo, and Bayu Satria Pratama, “KLASTERISASI DATA HASIL STUDI PELACAKAN TENTANG KARIR DAN PEKERJAAN LULUSAN PERGURUAN TINGGI MENGGUNAKAN ALGORITMA K-MEANS,” *J-Icon : Jurnal Komputer dan Informatika*, vol. 11, no. 2, pp. 157–164, Oct. 2023.

[4] A. BASTIAN, “Penerapan Algoritma K-Means Clustering Analysis Pada Penyakit Menular Manusia (Studi Kasus Kabupaten Majalengka),” *Jurnal Sistem Informasi*, vol. 14, no. 1, pp. 28–34, Apr. 2018, doi: 10.21609/jsi.v14i1.566.

[5] “Tracer Study,” Tracer Study POLBAN . Accessed: Oct. 29, 2023. [Online]. Available: https://penelusuranalumni.polban.ac.id/tentang#:~:text=Tracer%20Study%20atau%20yang%20sering,lulusan%20lembaga%20penyelenggara%20pendidikan%20tinggi

[6] C. Zai, “IMPLEMENTASI DATA MINING SEBAGAI PENGOLAHAN DATA,” *JURNAL PORTAL DATA*, vol. 2, Apr. 2022.

[7] P. S. Hasugian, J. R. Sagala, and L. D. Ani, “Alumni Data Grouping Using the K-Means Clustering Method for Study Program Curriculum Development,” *Jurnal Info Sains : Informatika dan Sains*, vol. 13, no. 2, pp. 137–144, 2023.

[8] A. Nofiar, S. Defit, and Sumijan, “Penentuan Mutu Kelapa Sawit Menggunakan Metode K-Means Clustering,” *Jurnal KomtekInfo*, vol. 5, no. 3, pp. 1–9, Apr. 2019, doi: 10.35134/komtekinfo.v5i3.26.

[9] A. Salam, D. Adiatma, and J. Zeniarja, “Implementasi Algoritma K-Means Dalam Pengklasteran untuk Rekomendasi Penerima Beasiswa PPA di UDINUS,” *JOINS (Journal of Information System)*, vol. 5, no. 1, pp. 62–68, May 2020, doi: 10.33633/joins.v5i1.3350.

[10] B. Orleans and E. P. Putra, “Clustering Algoritma (K-Means),” BINUS Higher Education.

[11] F. Febriansyah, “PENERAPAN ALGORITMA K-MEANS CLUSTERING DATA GIZI BALITA PADA UPTD PUSKESMAS BUMI AGUNG,” *Jurnal*

Informatika dan Teknik Elektro Terapan, vol. 12, no. 3, Aug. 2024, doi: 10.23960/jitet.v12i3.4923.

[12] I. W. Sukerta Wijaya, I. G. Harjumawan Wiratmaja KS., I. D. M. A. Pramana Setya Bintara, and I. K. G. Ryan Aditya Permana, "Program Menghitung Banyak Bata pada Ruangan Menggunakan Bahasa Python," *TIERS Information Technology Journal*, vol. 2, no. 1, Dec. 2021, doi: 10.38043/tiers.v2i1.2840.

[13] M Riziq Sirfatullah Alfarizi, Muhamad Zidan Al-farish, Muhamad Taufiqurrahman, Ginan Ardiansah, and Muhamad Elgar, "Penggunaan Python Sebagai Bahasa Pemrograman untuk Machine Learning dan Deep Learning," *Karimah Tauhid*, vol. 2, no. 1, pp. 1–6, Jan. 2023.

[14] S. Mahallati, J. C. Bezdek, M. R. Popovic, and T. A. Valiante, "Cluster tendency assessment in neuronal spike data," *PLoS One*, vol. 14, no. 11, p. e0224547, Nov. 2019, doi: 10.1371/journal.pone.0224547.

[15] A. Akiode, "Using Visualization Algorithms (VAT & iVAT) To Assess Cluster Tendency," Analytics Vidhya. Accessed: Feb. 25, 2024. [Online]. Available: https://medium.com/analytics-vidhya/using-visualization-algorithms-vat-ivat-to-assess-cluster-tendency-a89251a2400e

[16] D. Kumar and J. C. Bezdek, "Clustering tendency assessment for datasets having inter-cluster density variations," in *2020 International Conference on Signal Processing and Communications (SPCOM)*, IEEE, Jul. 2020, pp. 1–5. doi: 10.1109/SPCOM50965.2020.9179608.

[17] A. W. Fuadah, F. N. Arifin, and O. Juwita, "Optimasi K-Klasterisasi Ketahanan Pangan Kabupaten Jember Menggunakan Metode Elbow," *INFORMAL: Informatics Journal*, vol. 6, no. 3, p. 136, Dec. 2021, doi: 10.19184/isj.v6i3.28363.

[18] R. Nainggolan, R. Perangin-angin, E. Simarmata, and A. F. Tarigan, "Improved the Performance of the K-Means Cluster Using the Sum of Squared Error (SSE) optimized by using the Elbow Method," *J Phys Conf Ser*, vol. 1361, no. 1, p. 012015, Nov. 2019, doi: 10.1088/1742-6596/1361/1/012015.

[19] A. Fikri, B. F. Hutabarat, and U. Khaira, "Komparasi Antara Metode K-Means Clustering Dan Complete Linkage Dalam Pengelompokan Penyaluran Pinjaman Oleh Financial Technology," *Jurnal Ilmiah Media Sisfo*, vol. 17, no. 2, pp. 228–239, Oct. 2023, doi: 10.33998/mediasisfo.2023.17.2.1373.

[20] A. Bhardwaj, "Silhouette Coefficient Validating clustering techniques," Towards Data Science.

[21] Septian Wulandari, "Clustering Kecamatan Di Kota Bandung Berdasarkan Indikator Jumlah Penduduk Dengan Menggunakan Algoritma K-Means," *Seminar Nasional Riset dan Teknologi (SEMNAS RISTEK)*, vol. 4, no. 1, pp. 128–132, Jan. 2020.

[22] I. Permana and F. N. S. Salisah, "Pengaruh Normalisasi Data Terhadap Performa Hasil Klasifikasi Algoritma Backpropagation," *Indonesian Journal of Informatic Research and Software Engineering (IJIRSE)*, vol. 2, no. 1, pp. 67–72, Mar. 2022, doi: 10.57152/ijirse.v2i1.311.

[23] Hendro Priyatman, Fahmi Sajid, and Dannis Haldivany, "Klasterisasi Menggunakan Algoritma K-Means Clustering untuk Memprediksi Waktu Kelulusan Mahasiswa," *JEPIN (Jurnal Edukasi dan Penelitian Informatika)*, vol. 5, no. 1, pp. 62–66, Mar. 2023.

[24] K. R. Shahapure and C. Nicholas, "Cluster Quality Analysis Using Silhouette Score," in *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, IEEE, Oct. 2020, pp. 747–748. doi: 10.1109/DSAA49011.2020.00096.