Nama : Hajir Rizky Nugroho

NIM : A11.2019.12297

Kelompok : A11.4619

LAPORAN HASIL EKSPERIMEN TUGAS AKHIR DATA MINING

1. JUDUL

"Penerapan Metode Clustering K-Means untuk Segmentasi Customer Mall"

2. PEMBUATAN KODING PYTHON

Pada pembuatan koding untuk eksperimen data mining ini menggunakan bahasa pemrograman Python dan text editor Jupyter Notebook. Berikut adalah penjelasan mengenai langkah-langkah serta source code dalam eksperimen data mining tersebut :

- Import library

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

Library yang digunakan dalam eksperimen ini menggunakan pandas, numpy, matplotlib.

- Memanggil dataset dari file .csv

```
df = pd.read_csv("Mall_Customers.csv")
```

- Menampilkan data teratas

```
df.head(5)
```

- Me-rename kolom genre menjadi gender

	CustomerID	Genre	Age	Annual Income (k$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40

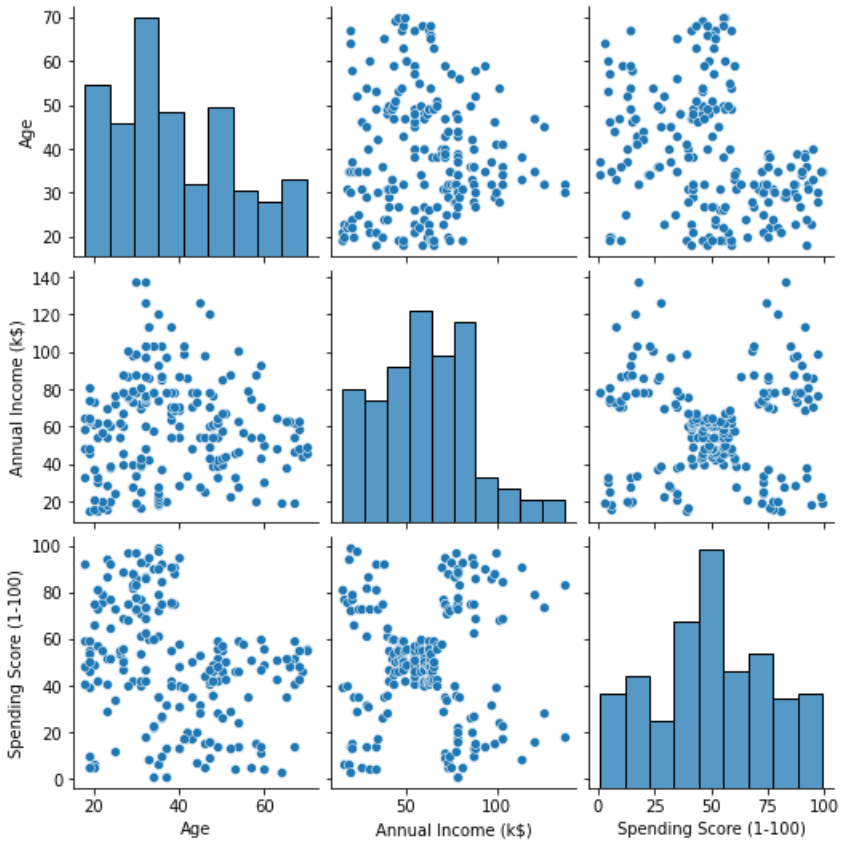
```
df = df.rename(columns={'Genre':'Gender'})
df.head(5) # Melihat kembali kolom yang di-rename
```

	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40

3. IMPLEMENTASI SOURCE CODE UNTUK KLASTERISASI

- Menghapus redudansi/preprocessing data

```
df.shape
# Mencari apakah ada value yang hilang
df.isnull().sum()

df.describe()
feature = df.drop(columns='CustomerID') #menghilangkan kolom Customer ID
sns.pairplot(feature,kind="scatter")
```

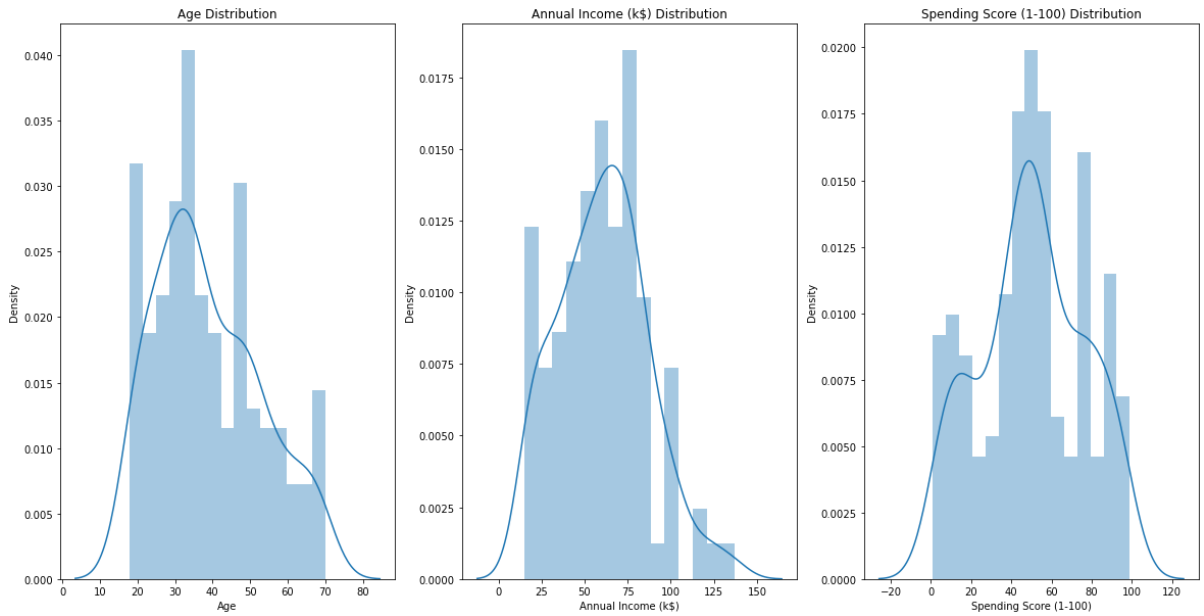


Jika kita melihat pada pairplot Annual Income dan Spending score, jelas bahwa customer dapat dibagi menjadi 5.
Menggabungkan Age dan spending, dapat dibagi menjadi 2 kelompok: kiri atas dan kanan bawah, daya beli kaum muda relatif lebih tinggi.

- Melakukan scalling data

```
# Rasio pria dan wanita
gender_count = df['Gender'].value_counts()
plt.pie(gender_count.values, labels=gender_count.index, labeldistance=1.2,
        wedgeprops = { 'linewidth' : 3, 'edgecolor' : 'white' })
```


```
plt.figure(figsize = (20 , 10))
n=1
for x in ['Age' , 'Annual Income (k$)' , 'Spending Score (1-100)']:
    plt.subplot(1,3,n)
    plt.subplots_adjust(hspace =0.5 , wspace = 0.2)
    sns.distplot(df[x], bins = 15)
    plt.title('{} Distribution'.format(x))
    n+=1
```



Jika kita menyelam lebih dalam ke fitur, diamati bahwa spending score memiliki 3 puncak (0-20,40-60,80-100), sedangkan untuk Annual Income, cenderung miring ke kanan.

4. **Klastering dengan k-means**

```
from sklearn.cluster import KMeans
```

K-means clustering merupakan salah satu jenis unsupervised learning, biasanya digunakan ketika kita tidak mengetahui kelompok/kategori mereka. Algoritme menetapkan setiap titik data ke salah satu grup K berdasarkan kesamaan fitur.

Hal ini berguna untuk menemukan kelompok yang belum secara eksplisit diberi label dalam data. Ini dapat digunakan untuk mengkonfirmasi asumsi bisnis tentang jenis grup yang ada atau untuk mengidentifikasi grup yang tidak dikenal dalam kumpulan data yang kompleks.

```
# Mengubah fitur non-numerik menjadi angka
# Male=0, Female=1
feature.loc[feature['Gender']=='Male','Gender']=0
feature.loc[feature['Gender']=='Female','Gender']=1
```

Elbow Method

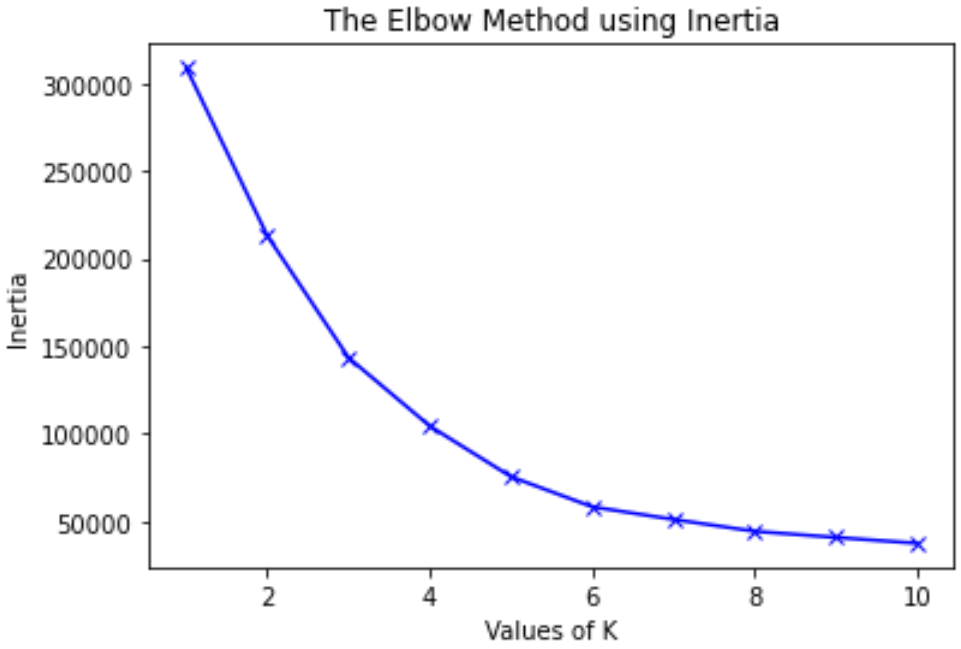
- Tujuan: Menemukan nilai optimal k di K-Means
- Pokok:
  memplot nilai fungsi biaya yang dihasilkan oleh nilai k yang berbeda. Ketika k meningkat, titik data dapat lebih lanjut "dipilih" ke cluster terdekat dan jarak antara masing-masing centroid akan berkurang. Namun, peningkatan sum of squared error (SSE) akan menurun dan mulai mendatar dengan meningkatnya k. Distorsi seperti itu menyerupai "elbow".

  Ketentuan:
  1. Distorsi rata-rata jarak kuadrat euclidean dari pusat massa masing-masing cluster
  2. Inersia Jumlah kuadrat jarak sampel ke pusat cluster terdekat

  Distorsi terutama muncul ketika k berada di antara 3 dan 5, jadi kita harus mencoba k=3/4/5.

```
inertias = []
for i in range(1, 11):
    km = KMeans(n_clusters=i).fit(feature)
    inertias.append(km.inertia_)
```

```
plt.plot(range(1, 11), inertias, 'bx-')
plt.xlabel('Values of K')
plt.ylabel('Inertia')
plt.title('The Elbow Method using Inertia')
plt.show()
```

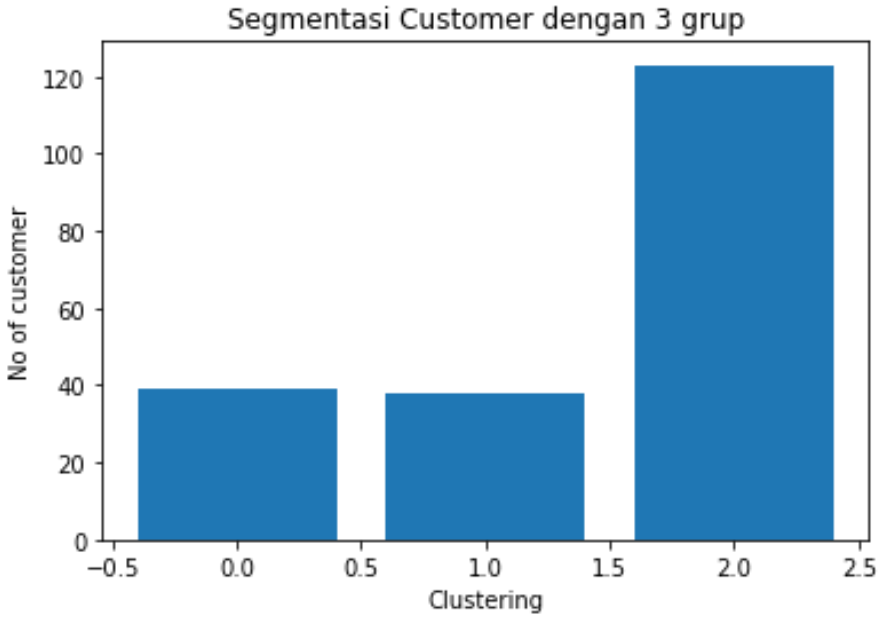



```
inertias
```

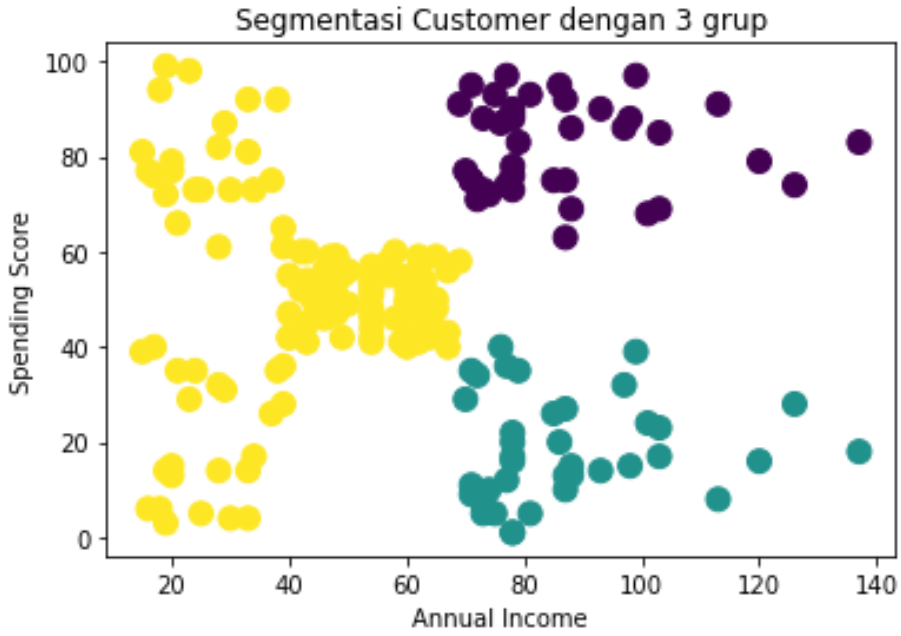
5. **Modelling dengan metode K-means**

```
# 3 klastering
# Jumlah customer di setiap grup masing-masing
km = KMeans(n_clusters=3).fit(feature)
y_km = km.fit_predict(feature)
n_cluster, km_count = np.unique(y_km, return_counts=True)
plt.bar(n_cluster, km_count)
plt.ylabel('No of customer')
plt.xlabel('Clustering')
```

```
plt.title('Segmentasi Customer dengan 3 grup')
```

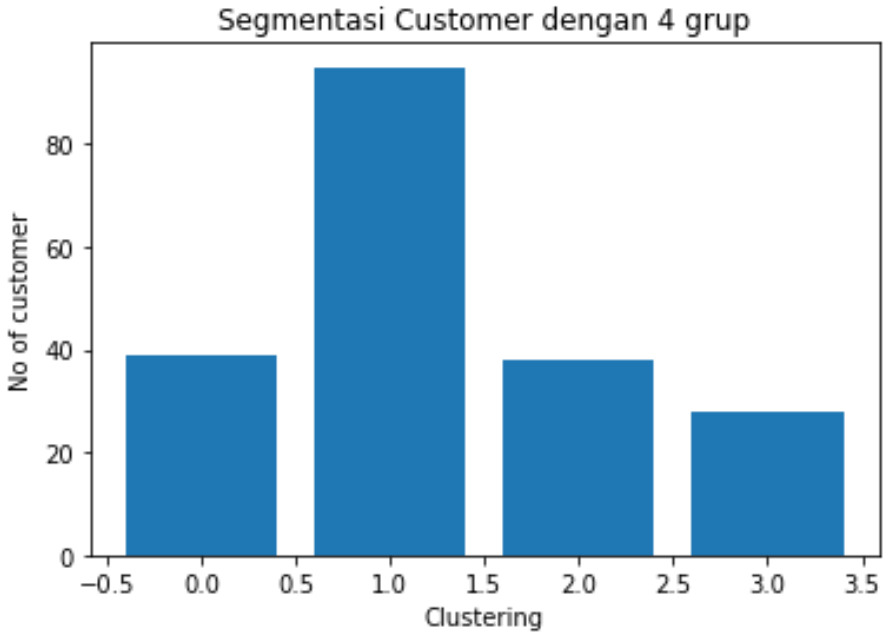



```
plt.scatter(df['Annual Income (k$)'],
            df['Spending Score (1-100)'],
            c=y_km, s=100)


plt.xlabel('Annual Income')
plt.ylabel('Spending Score')
plt.title('Segmentasi Customer dengan 3 grup')

plt.show()
```



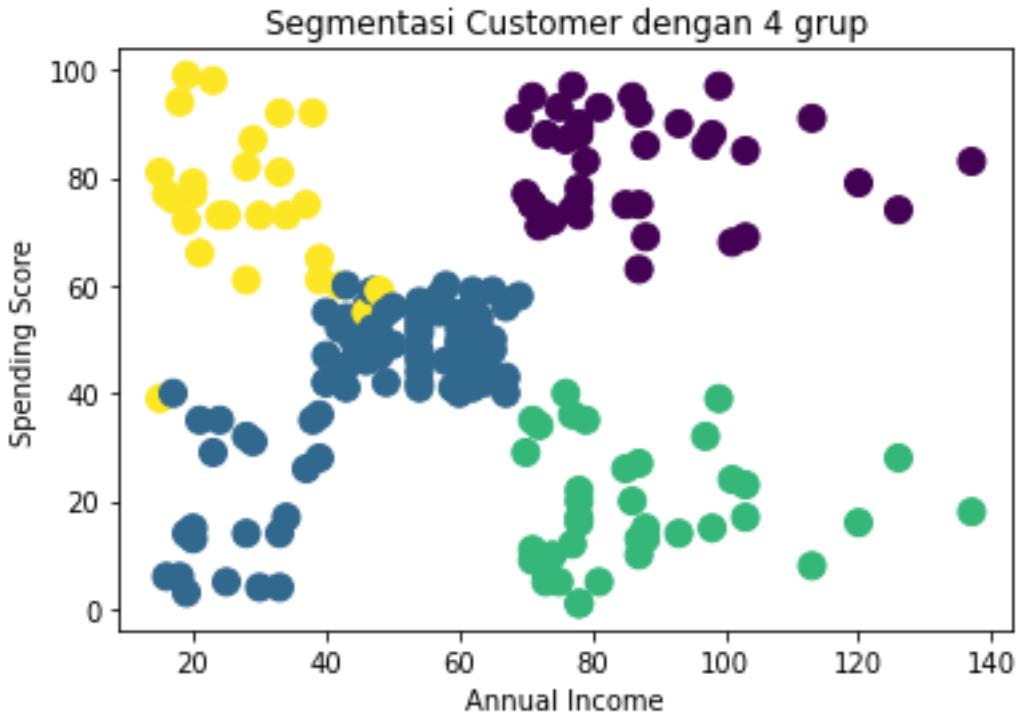

```
# 4 Klastering
# Jumlah customer di setiap grup masing-masing
km = KMeans(n_clusters=4).fit(feature)
```

```
y_km = km.fit_predict(feature)
n_cluster, km_count = np.unique(y_km, return_counts=True)
plt.bar(n_cluster, km_count)
plt.ylabel('No of customer')
plt.xlabel('Clustering')
plt.title('Segmentasi Customer dengan 4 grup')
plt.scatter(df['Annual Income (k$)'],
            df['Spending Score (1-100)'],
            c=y_km, s=100)


plt.xlabel('Annual Income')
plt.ylabel('Spending Score')
plt.title('Segmentasi Customer dengan 4 grup')

plt.show()
```



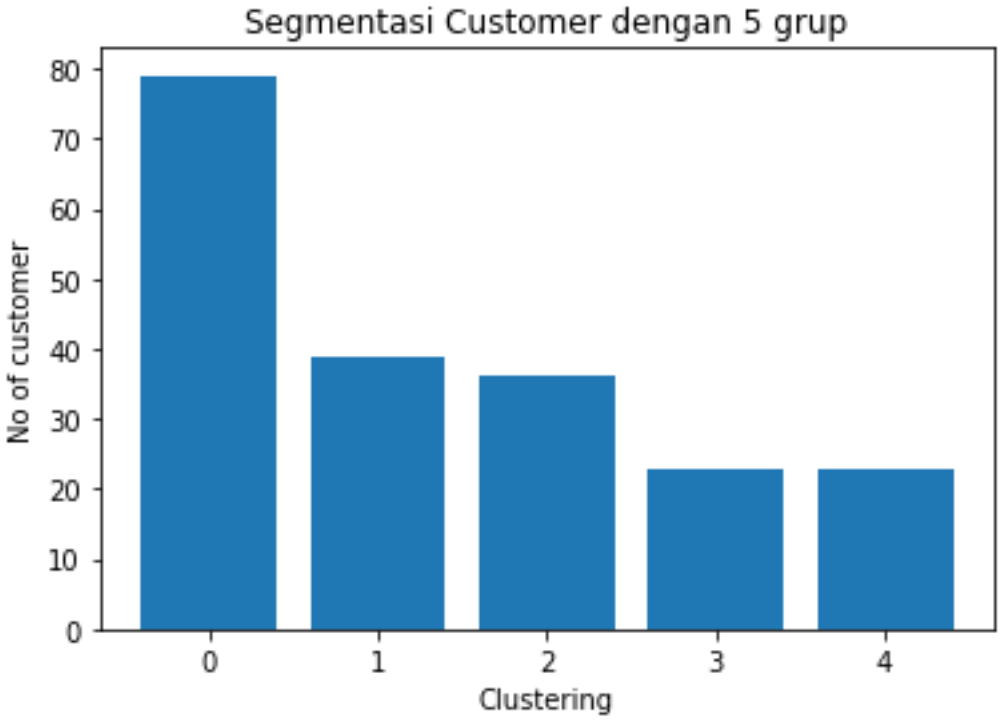

```
plt.scatter(df['Annual Income (k$)'],
            df['Spending Score (1-100)'],
            c=y_km, s=100)


plt.xlabel('Annual Income')
plt.ylabel('Spending Score')
plt.title('Segmentasi Customer dengan 4 grup')
```

```
plt.show()
```

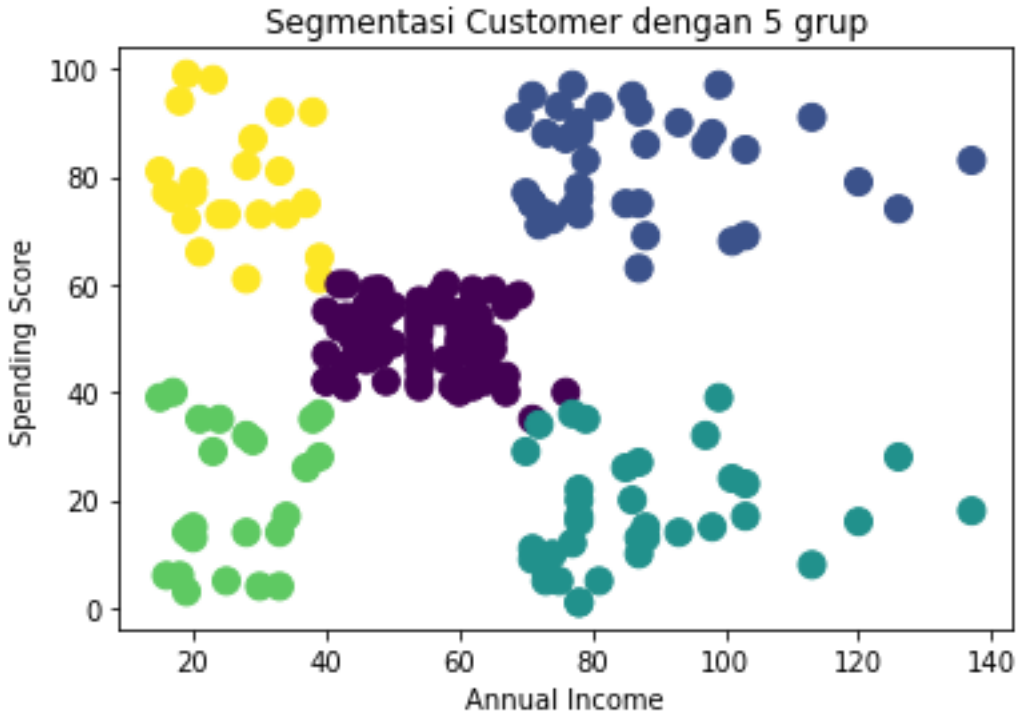



```
# 5 Klastering
# Jumlah customer di setiap grup masing-masing
km = KMeans(n_clusters=5).fit(feature)
y_km = km.fit_predict(feature)
n_cluster, km_count = np.unique(y_km, return_counts=True)
plt.bar(n_cluster, km_count)
plt.ylabel('No of customer')
plt.xlabel('Clustering')
```

```
plt.title('Segmentasi Customer dengan 5 grup')
```




```
plt.scatter(df['Annual Income (k$)'],
            df['Spending Score (1-100)'],
            c=y_km, s=100)


plt.xlabel('Annual Income')
plt.ylabel('Spending Score')
plt.title('Segmentasi Customer dengan 5 grup')

plt.show()
```



6. **Hasil Kesimpulan**

Dari eksperimen yang dilakukan, dapat diketahui daya beli kaum muda relatif lebih tinggi. Dan didapatkan hasil klasterisasi dengan 5 grup yaitu hijau, kuning, ungu, biru tua, dan biru muda.

Grup hijau memiliki pendapatan yang rendah namun dengan pengeluaran yang rendah juga. Grup kuning memiliki yang rendah tetapi memiliki pengeluaran yang cukup tinggi. Grup ungu memiliki pendapatan yang cukup dan pengeluaran yang cukup juga. Grup biru tua memiliki pendapatan yang tinggi dan juga pengeluaran yang tinggi. Grup biru muda memiliki pendapatan yang tinggi namun memiliki pengeluaran yang rendah.

RINGKASAN EKSPERIMEN TUGAS AKHIR DATA MINING

1. JUDUL

"Penerapan Metode Clustering K-Means untuk Segmentasi Customer Mall"

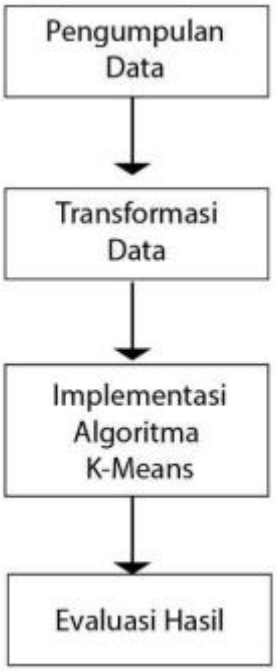
2. LATAR BELAKANG MASALAH

Pada perkembangan zaman saat ini, banyak kemudahan yang dapat dilakukan masyarakat melalui teknologi dan banyak pengaruhnya. Salah satunya adalah tindakan konsumtif masyarakat, mulai dari kalangan muda hingga lanjut usia. Hal ini lah yang disukai oleh banyak perusahan. Banyak pemilik perusahaan, khususnya mall yang memanfaatkan hal ini.

Setiap customer mall memiliki umur dan pendapatan yang berbeda-beda namun beda juga tindakan konsumtif mereka. Dalam kasus ini, segmentasi customer mall diperlukan untuk perusahaan agar memaksimalkan kesempatan yang didapatkan dari hal ini. Untuk mengkelompokkan customer dapat menggunakan metode klasterisasi dengan k-means. K-means clustering merupakan salah satu jenis unsupervised learning, biasanya digunakan ketika kita tidak mengetahui kelompok/kategori mereka. Algoritme menetapkan setiap titik data ke salah satu grup K berdasarkan kesamaan fitur. Hal ini berguna untuk menemukan kelompok yang belum secara eksplisit diberi label dalam data. Ini dapat digunakan untuk mengkonfirmasi asumsi bisnis tentang jenis grup yang ada atau untuk mengidentifikasi grup yang tidak dikenal dalam kumpulan data yang kompleks.

3. METODE PENYELESAIAN

Penelitian ini dilakuan dengan melewati empat tahapan diantaranya : pengumpulan data, transformasi data, implementasi algoritma k-means dan evaluasi hasil.



4. TAHAPAN PENELITIAN

- Menentukan dataset yang digunakan yang berisi umur,pendapatan dan ukuran pengeluaran.
- Melakukan preprocessing dan edit data pada dataset agar siap digunakan sesuai kebutuhan
- Pembuatan koding python
- Pengujian dengan menggunakan metode Clustering K-Means
- Mengevaluasi hasil penelitian

5. Kesimpulan

Dari eksperimen yang dilakukan, dapat diketahui daya beli kaum muda relatif lebih tinggi. Dan didapatkan hasil klasterisasi dengan 5 grup yaitu hijau, kuning, ungu, biru tua, dan biru muda.

Grup hijau memiliki pendapatan yang rendah namun dengan pengeluaran yang rendah juga. Grup kuning memiliki yang rendah tetapi memiliki pengeluaran yang cukup tinggi. Grup ungu memiliki pendapatan yang cukup dan pengeluaran yang cukup juga. Grup biru tua memiliki pendapatan yang tinggi dan juga pengeluaran yang tinggi. Grup biru muda memiliki pendapatan yang tinggi namun memiliki pengeluaran yang rendah.

DAFTAR PUSTAKA

Hakim, L., & Seruni, H. (2018). Indikasi Penyimpangan Laporan Keuangan Akademik Universitas XYZ Menggunakan Algoritma Greedy dan K-Means. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *2*(1), 301–306. https://doi.org/10.29207/resti.v2i1.261

K-means, M. M., Azis, S. A., Defit, S., & Yunus, Y. (2021). *Jurnal Informatika Ekonomi Bisnis Klasterisasi Dana Bantuan Pada Program Keluarga Harapan (PKH)*. *3*, 53–59. https://doi.org/10.37034/infeb.v3i2.66

Rahayu, A. E., Hikmah, K., Yustia, N., & Fauzan, A. C. (2019). Penerapan K-Means Clustering Untuk Penentuan Klasterisasi Beasiswa Bidikmisi Mahasiswa. *ILKOMNIKA: Journal of Computer Science and Applied Informatics*, *1*(2), 82–86. https://doi.org/10.28926/ilkomnika.v1i2.23

Sinaga, K. P., & Yang, M. S. (2020). Unsupervised K-means clustering algorithm. *IEEE Access*, *8*, 80716–80727. https://doi.org/10.1109/ACCESS.2020.2988796

Siti, M. (2018). Segmentasi Perilaku Pembelian Pelanggan Berdasarkan Model RFM dengan Metode K-Means. *Jurnal Sistem Informasi*, *2*(1), 9–15.