KLASTERING DATA PASIEN STROKE MENGGUNAKAN METODE K-MEANS

Disusun Oleh :

Shaka Rizky Ramadhan
A11.2019.12214

Program Studi Teknik Informatika
Fakultas Ilmu Komputer
Universitas Dian Nuswantoro

BAB I
TINJAUAN PUSTAKA DAN LANDASAN TEORI

1.1 Tinjauan Studi

Sebuah penelitian pasti selalu berhubungan dengan penelitian sebelumnya, penelitian sebelumnya inilah yang menjadikan sebuah penelitian mempunyai gambaran dan dapat dijadikan sebagai batasan serta langkah awal untuk lebih baik dari penelitian sebelumnya.

Berikut adalah beberapa penelitian sebelumnya yang memiliki tema yang relevan dengan penelitian ini dan didapatakn dari sumber yang kredibel.

Tabel 1. *State of Art*

No	Peneliti	Tahun	Judul	Metode	Hasil Penelitian
1	Fitri Larasati Sibuea, Andy Sapta	2017	Pemetaan Siswa Berprestasi Menggunakan Metode K-Means Clustering	K-Means	Metode k-means clustering dapat membantu mengelompokkan siswa dengan prestasi tinggi, menengah dan cukup.
2	Gustientiedina, M.Hasmil Adiya, Yenny Desnelita	2019	Penerapan Algoritma K-Means Untuk Clustering Data Obat-Obatan Pada RSUD Pekanbaru	K-Means	Dengan menggunakan algoritma k-means untuk klastering, RSUD Pekanbaru dapat mengelompokkan obat berdasarkan banyaknya permintaan setiap tahunnya.
3	Hendro Priyatman, Fahmi Sajid, Dannis Haldivanny	2019	Klasterisasi Menggunakan Algoritma K-Means Untuk Memprediksi Waktu Kelulusan Mahasiswa	K-Means	Setelah dilakukan pengujian dan analisis program maka dapat diperoleh kesimpulan bahwa pada kasus ini implementasi algoritma k-means dalam data mining sudah berhasil, dan bisa menampilkan informasi prediksi kelulusan

					mahasiswa, namun untuk melihat tingkat kemampuan real k-means clustering dalam memprediksi waktu kelulusan tergantung pada mahasiswa itu sendiri
3	Suhandio Handoko, Fauziah, Endah Tri Esti Handayani	2020	Implementasi Data Mining Untuk Menentukan Tingkat Penjualan Paket Data Telkomsel Menggunakan Metode K-Means Clustering	K-Means	Algoritma K-Means Clustering data mining didapatkan daerah penjualan produk yang tinggi , sedang , dan rendah. Daerah dengan penjualan produk yang rendah akan dilakukan promosi penjualan produk dan untuk daerah penjualan yang tinggi tidak diadakan promosi.
4	Zulfa Nabila, Auliya Rahman Isnain, Permata, Zaenal Abidin	2021	Analisis Data Mining Untuk Clusteirng Kasus COVID-19 Di Provinsi Lampung Dengan Algoritma K-Means	K-Means	Berdasarkan hasil pengujian data menggunakan RapidMiner pada kasus Covid-19 di Provinsi Lampung per akhir bulan Maret 2020 sampai dengan Maret 2021 menghasilkan pengelompokkan yang berbeda

					dikarenakan jumlah pada atribut Suspek, Probable, Konfirmasi Positif, Selesai Isolasi, dan Kematian pada setiap Kabupaten/Kota yang tidak sama.

1.2 Landasan Teori

- **K-Means**

  K-means merupakan salah satu algoritma yang bersifat unsupervised learning. K-Means memiliki fungsi untuk mengelompokkan data kedalam data cluster. Algoritma ini dapat menerima data tanpa ada label kategori. K-Means Clustering Algoritma juga merupakan metode non-hierarchy. Metode Clustering Algoritma adalah mengelompokkan beberapa data ke dalam kelompok yang menjelaskan data dalam satu kelompok memiliki karakteristik yang sama dan memiliki karakteristik yang berbeda dengan data yang ada di kelompok lain. Cluster Sampling adalah teknik pengambilan sampel di mana unit-unit populasi dipilih secara acak dari kelompok yang sudah ada yang disebut 'cluster, nah Clustering atau klasterisasi adalah salah satu masalah yang menggunakan teknik *unsupervised learning*.

- **Kaggle**

  Kaggle adalah sebuah komunitas *online* yang dibentuk oleh Anthony Goldbloom sebagai CEO dan Ben Hamner sebagai CTO di tahun 2010. Komunitas *online* ini menampung para pegiat *data science* yang ingin belajar lebih dalam tentang *machine learning* dan ilmu-ilmu terkait lainnya. Di dalamnya, ada berbagai kegiatan yang dilakukan, salah satunya yang paling terkenal adalah kompetisi *machine learning.* Kaggle sendiri menyatakan bahwa selain berkompetisi, anggota komunitasnya bisa bersama-sama menulis dan membagikan kode serta mempelajari berbagai hal. Bahkan, para *data scientist* bisa juga mendapatkan uang dari proyek yang ditawarkan di Kaggle, lho. Saat ini, Kaggle telah memiliki lebih dari 1000 *dataset*, 170.000 *post* di forum, dan paling tidak 250 *kernel*.

- **Google Collab**

  Google Colab atau Google Colaboratory, adalah sebuah *executable document* yang dapat digunakan untuk menyimpan, menulis, serta membagikan program yang telah ditulis melalui Google Drive. *Software* ini pada dasarnya serupa dengan Jupyter Notebook gratis berbentuk *cloud* yang dijalankan menggunakan *browser*, seperti Mozilla Firefox dan Google Chrome. Ia memungkinkan penggunanya untuk menjalankan kode Python tanpa perlu melakukan proses instalasi dan *setup* lainnya. Justru, semua keperluan *setting* dan *adjustment* akan diserahkan ke *cloud*.

- **Python**
  Python adalah bahasa pemrograman interpretatif multiguna. Tidak seperti bahasa lain yang susah untuk dibaca dan dipahami, python lebih menekankan pada keterbacaan kode agar lebih mudah untuk memahami sintaks. Hal ini membuat Python sangat mudah dipelajari baik untuk pemula maupun untuk yang sudah menguasai bahasa pemrograman lain. Bahasa ini muncul pertama kali pada tahun 1991, dirancang oleh seorang bernama Guido van Rossum. Sampai saat ini Python masih dikembangkan oleh Python Software Foundation. Bahasa Python mendukung hampir semua sistem operasi, bahkan untuk sistem operasi Linux, hampir semua distronya sudah menyertakan Python di dalamnya.

BAB II
PEMBAHASAN

2.1 Dataset

Dataset yang digunakan kali ini didapatkan dari Kaggle yang berupa data pasien dengan umur dan rata-rata gula darah yang dimiliki. Dataset ini terdiri dari 150 data. Nantinya data ini yang akan digunakan sebagai dataset untuk klasterisasi pasien stroke. Berikut adalah contoh dataset yang digunakan.

	Age	AVG Glucose Level
0	67	228.69
1	61	202.21
2	80	105.92
3	49	171.23
4	79	174.12

2.2 Proses Klasterisasi

Langkah pertama yang akan kita lakukan yaitu melakukan import library python yang akan digunakan untuk klasterisasi menggunakan algoritma k-means. Kita bisa melakukan proses klasterisasi menggunakan Google Collab yang bisa diakses melalui web browser.

```
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
from sklearn.cluster import KMeans
```

Seperti yang ada pada gambar diatas, beberapa library yang digunakan yaitu matplotlib yang berfungsi untuk memvisualisasikan data seperti membuat plot grafik untuk satu sumbu atau lebih. Selanjutnya library numpy digunakan berfungsi untuk melakukan operasi vector dan matriks dengan mengolah array dan array multidimensi. Untuk library pandas berfungsi sebagai mengelola data berbentuk table. Yang terakhir KMeans sendiri library yang berisi algoritma K-Means untuk melakukan klasterisasi.

Selanjutnya adalah melakukan import dataset yang telah didapatkan sebelumnya dari Kaggle. Dataset yang digunakan dalam bentuk file .csv dengan dua kolom yaitu “Age” dan “AVG Glucose Level”.

```
[2] #Import dataset
    dataset = pd.read_csv('stroke.csv', sep=';')
    dataset.keys()

    Index(['Age', ' AVG Glucose Level'], dtype='object')
```

Setelah dataset di import, Langkah selanjutnya adalah memastikan apakah dataset tersebut bisa terbaca dan sesuai dengan dataset yang dimiliki. Untuk itu kita akan menampilkan 5 data teratas dari file dataset tersebut.

```
#Memanggil 5 data pertama
myData = pd.DataFrame(dataset)
myData.head()
```

	Age	AVG Glucose Level
0	67	228.69
1	61	202.21
2	80	105.92
3	49	171.23
4	79	174.12

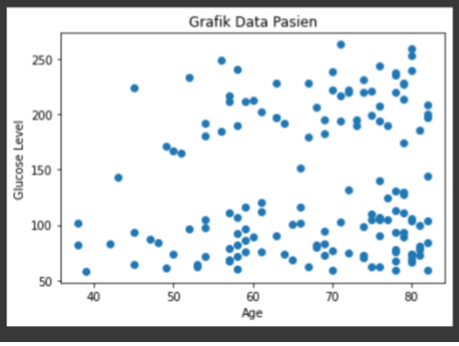
Apabila data sudah terbaca dan sesuai dengan dataset yang kita miliki, Langkah selanjutnya adalah memanggil data tersebut dan memasukkannya dalam bentuk array menggunakan library numpy yang sudah kita import sebelumnya.

```
#Memanggil data dalam bentuk array
X = np.asarray(dataset)
print(X)

[[ 67.   228.69]
 [ 61.   202.21]
 [ 80.   105.92]
 [ 49.   171.23]
 [ 79.   174.12]
 [ 81.   186.21]
 [ 74.    70.09]
 [ 69.    94.39]
 [ 59.    76.15]
 [ 78.    58.57]
 [ 81.    80.43]
 [ 61.   120.46]
```

Langkah selanjutnya adalah menampilkan data dalam bentuk plot grafik menggunakan library matplotlib yang sudah kita import sebelumnya. Dalam plot grafik ini memiliki dua sumbu yaitu “Age” sebagai sumbu X dan “AVG Glucose Level” sebagai sumbu Y.

```
#Menampilkan data dalam bentuk scatter plot
plt.scatter(X[:,0], X[:,1], label='True Position')
plt.xlabel("Age")
plt.ylabel("Glucose Level")
plt.title("Grafik Data Pasien")
plt.show()
```



Setelah itu kita akan menentukan jumlah klaster yang akan dibuat. Untuk kasus ini, jumlah klaster yang akan dibuat yaitu berjumlah dua klaster. Pembuatan klaster ini menggunakan library KMeans dari skcit.learn yang sudah di import sebelumnya.

```
#Mengaktifkan KMeans dengan jumlah k=2
kmeans = KMeans(n_clusters=2)
kmeans.fit(X)
```

```
KMeans(n_clusters=2)
```

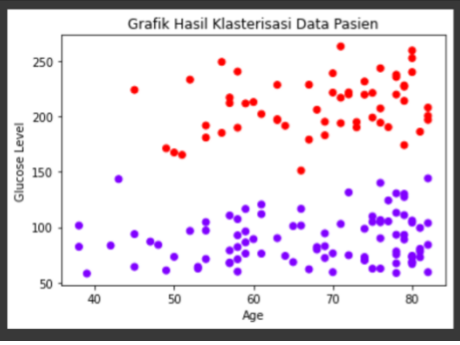
Setelah menentukan jumlah klaster yang akan dibuat, Langkah selanjutnya adalah menentukan nilai centroid yang akan digunakan sebagai nilai titik tengah dalam suatu klaster. Penentuan nilai centroid ini juga menggunakan library KMeans yang sudah di import sebelumnya.

```
[ ] #Menampilkan nilai centroid yang telah dibuat oleh algoritma
    print(kmeans.cluster_centers_)
```

```
[[ 67.12903226  88.71569892]
 [ 68.40350877 209.8654386 ]]
```

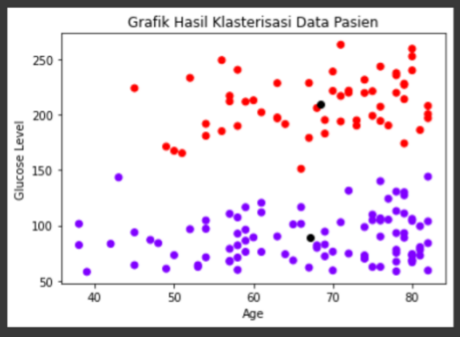
Seperti yang ada pada gambar diatas, nilai centroid yang didapatkan yaitu [67.123, 88.716] untuk klaster pertama dan [68.403, 209.865] untul klaster kedua. Untuk Langkah selanjutnya yaitu menampilkan data yang sudah dikelompokkan dalam bentuk plot grafik dengan library matplotlib.

```
#Plot data point
#Memvisualisasikan hasil klasterisasi data pasien
plt.scatter(X[:,0], X[:,1], c=kmeans.labels_, cmap='rainbow')
plt.xlabel("Age")
plt.ylabel("Glucose Level")
plt.title("Grafik Hasil Klasterisasi Data Pasien")
plt.show()
```



Seperti yang ada pada gambar diatas, titik-titik pada grafik tersebut dibagi menjadi dua warna yaitu merah dan biru. Warna merah berarti klaster pertama dan warna biru untuk klaster kedua. Setelah itu, Langkah selanjutnya adalah menampilkan nilai centroid dalam bentuk titik pada grafik tersebut.

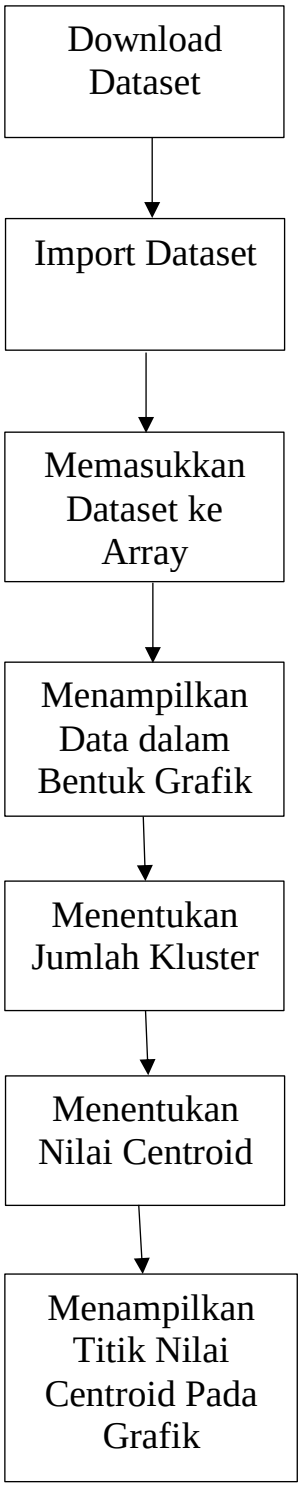
```
#Plot data point
#Memvisualisasikan hasil klasterisasi dengan centroid dari masing-masing kluster
plt.scatter(X[:,0], X[:,1], c=kmeans.labels_, cmap='rainbow')
plt.scatter(kmeans.cluster_centers_[:,0], kmeans.cluster_centers_[:,1], color='black')
plt.xlabel("Age")
plt.ylabel("Glucose Level")
plt.title("Grafik Hasil Klasterisasi Data Pasien")
plt.show()
```



Seperti yang terlihat pada gambar diatas, nilai centroid pada tiap klaster yang berwarna hitam berada pada titik sesuai dengan yang didapatkan di awal pada saat menentukan nilai centroid. Dengan begitu, berdasarkan dataset yang digunakan dikelompokkan menjadi dua klaster yang ternyata lebih banyak pada klaster kedua.

BAB III
TIMELINE DAN KENDALA

3.1 Timeline



3.2 Kendala

- Mencari dataset yang bisa digunakan untuk klastering. Karena data harus berupa nilai atau integer
- Menentukan nilai centroid tiap klaster

SARAN DAN KESIMPULAN

Klasterisasi yang diimplementasikan pada dataset pasien stroke berhasil dibuat dengan jumlah klaster 2 dan nilai centroid pada masing-masing klaster [67.123, 88.716] untuk klaster pertama dan [68.403, 209.865] untuk klaster kedua.

Saran untuk penelitian selanjutnya mungkin bisa ditentukan Kembali untuk nilai centroid berdasarkan data yang sudah terklasterisasi sehingga akan mendapatkan hasil yang lebih akurat.

DAFTAR PUSTAKA

Fitri Larasati Sibuea, A. S., 2017. Pemetaan Siswa Berprestasi Menggunakan Algoritma K-Means Clustering. *Jurnal Teknologi dan Sistem Informasi,* pp. 85-92.

Gustientiedina, M. A. Y. D., 2019. Penerapan Algoritma K-Means untuk Clustering Data Obat-Obat pada RSUD Pekanbaru. *Jurnal Nasional Teknologi dan Sistem Informasi,* 05(01), pp. 017-024.

Hendro Priyatman, F. S. D. H., 2019. Klasterisasi Menggunakan Algoritma K-Means Clustering untuk Mempredikasi Kelulusan Mahasiswa. *Jurnal Edukasi dan Penelitian Informatika,* 05(01), pp. 20-26.

Suhandio Handoko, F. E. T. E. H., 2020. IMPLEMENTASI DATA MINING UNTUK MENENTUKAN TINGKAT PENJUALAN PAKET DATA TELKOMSEL MENGGUNAKAN METODE K-MEANS CLUSTERING. *Jurnal Ilmiah Teknologi dan Rekayasa,* 25(01), pp. 76-88.

Zulfa Nabila, A. R. I. P. Z. A., 2021. ANALISIS DATA MINING UNTUK CLUSTERING KASUS COVID-19 DI PROVINSI LAMPUNG DENGAN ALGORITMA K-MEANS. *Jurnal Teknologi dan Sistem Informasi,* 2(1), pp. 100-108.