K-MEANS CLUSTERING
PRAKTIKUM PEMBELAJARAN MESIN

Disusun Oleh:

Nashirudin Baqiy 24060119130045

Lab A2

DEPARTEMEN ILMU KOMPUTER / INFORMATIKA

FAKULTAS SAINS DAN MATEMATIKA

UNIVERSITAS DIPONEGORO

2021

BAB I PENDAHULUAN

1.1. Rumusan Masalah

1.1.1. Lakukanlah *clustering* menggunakan dataset iris seperti yang digunakan pada praktikum sebelumnya!

1.1.2. Lakukan evaluasi hasil *clustering* menggunakan *inertia* dan *silhouette coefficient*!

1.2. Tujuan

1.2.1. Melakukan *clustering* menggunakan dataset iris.

1.2.2. Melakukan evaluasi hasil *clustering* dengan *inertia* dan *silhouette coefficient*.

1.3. Dasar Teori

Pembelajaran mesin (ML) adalah studi tentang algoritma komputer yang dapat ditingkatkan secara otomatis melalui pengalaman dan penggunaan data. Itu dilihat sebagai bagian dari kecerdasan buatan. Algoritma pembelajaran mesin membangun model berdasarkan data sampel, yang dikenal sebagai "data pelatihan", untuk membuat prediksi atau keputusan tanpa diprogram secara eksplisit untuk melakukannya. Algoritma pembelajaran mesin digunakan dalam berbagai macam aplikasi, seperti dalam kedokteran, penyaringan email, pengenalan suara, dan visi komputer, di mana sulit atau tidak mungkin untuk mengembangkan algoritma konvensional untuk melakukan tugas-tugas yang diperlukan.

Bagian dari pembelajaran mesin terkait erat dengan statistik komputasi, yang berfokus pada pembuatan prediksi menggunakan komputer; tetapi tidak semua pembelajaran mesin adalah pembelajaran statistik. Studi optimasi matematika memberikan metode, teori, dan domain aplikasi ke bidang pembelajaran mesin. Data mining adalah bidang studi terkait, dengan fokus pada analisis data eksplorasi melalui pembelajaran tanpa pengawasan. Beberapa implementasi pembelajaran mesin menggunakan data dan jaringan saraf dengan cara yang meniru kerja otak biologis. Dalam penerapannya di seluruh masalah bisnis, machine learning juga disebut sebagai analitik prediktif.

BAB II PEMBAHASAN

2.1. Deskripsi dataset

Dataset yang digunakan adalah dataset bunga iris yang sudah sangat terkenal.

2.2. Pemisahan fitur dan penghilangan kelas

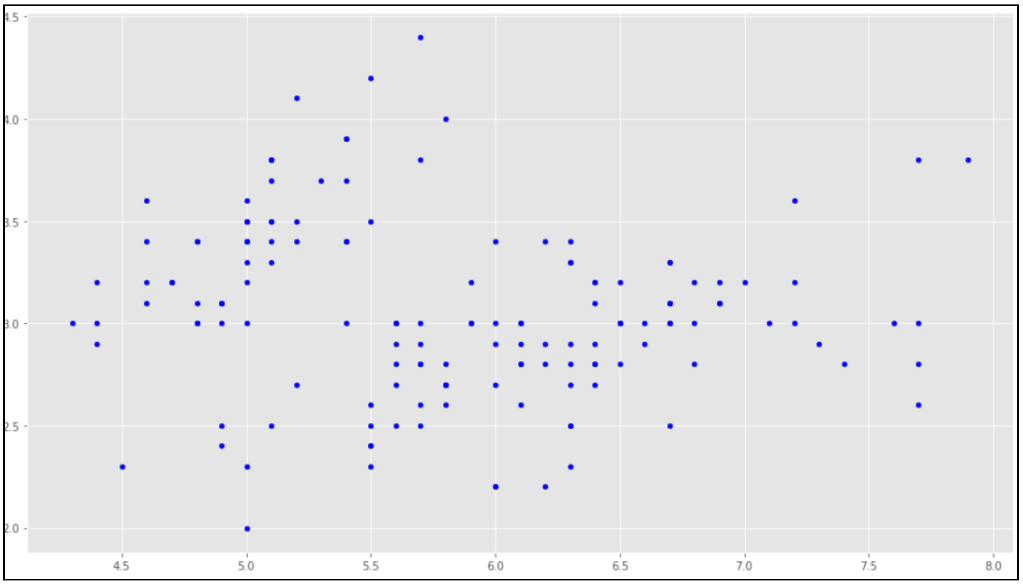
Karena dataset bunga iris ini bukan merupakan dataset yang tidak memiliki label melainkan sudah ada labelnya maka kita perlu melakukan proses untuk menghilangkan kelas dari dataset ini. Cara yang dilakukan di sini adalah membiarkan kolom kelas tidak digunakan sebagai anggota dalam proses *clustering* nantinya.

```
# Get the values of sepals
f1 = df['sepal-length'].values
f2 = df['sepal-width'].values
# print(f1)

# Get the values of petals
f3 = df['petal-length'].values
f4 = df['petal-width'].values
# print(f4)

X = np.array(list(zip(f1, f2)))
X2 = np.array(list(zip(f3, f4)))
print(X)
```

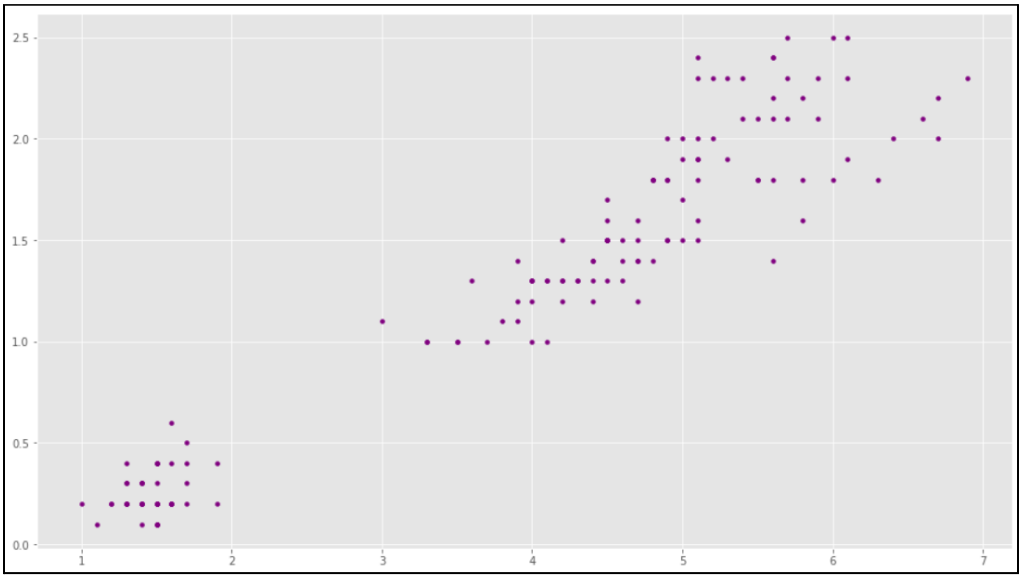
Kode di atas akan menghasilkan hasil sesuai dengan yang diinginkan, yaitu tidak adanya kolom kelas. Proses di atas juga akan membentuk dua kelompok klaster, X akan menghasilkan kluster sepal dan X2 akan menghasilkan petal.

2.3. Graph Plotting

Sepal:



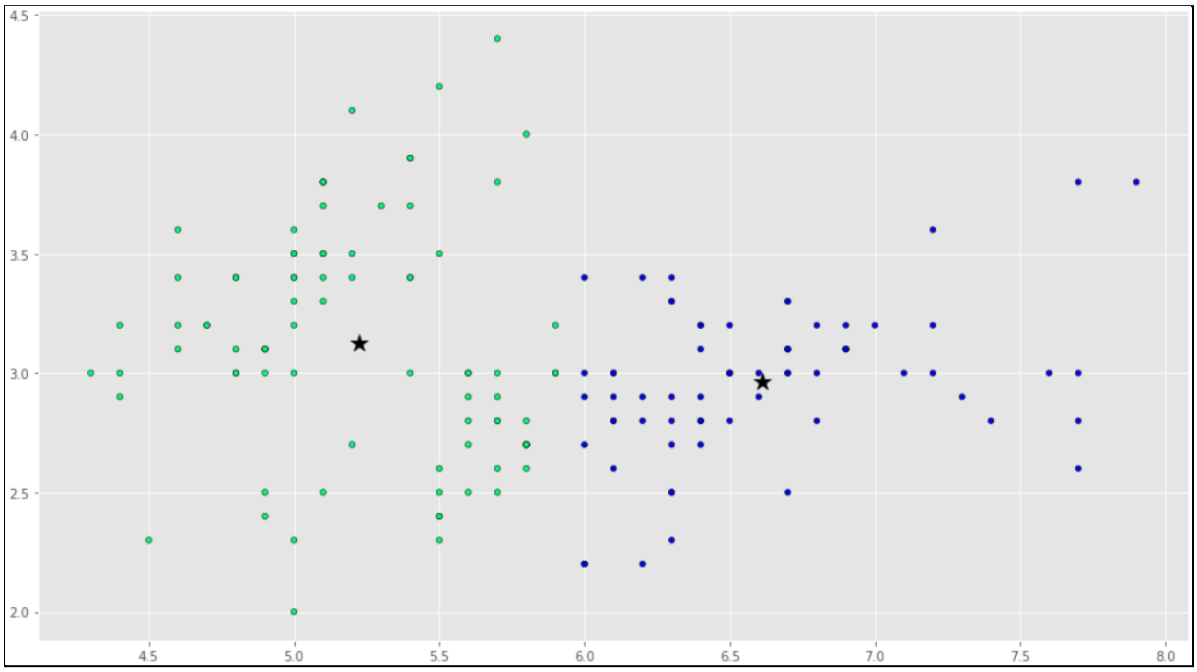
Petal:



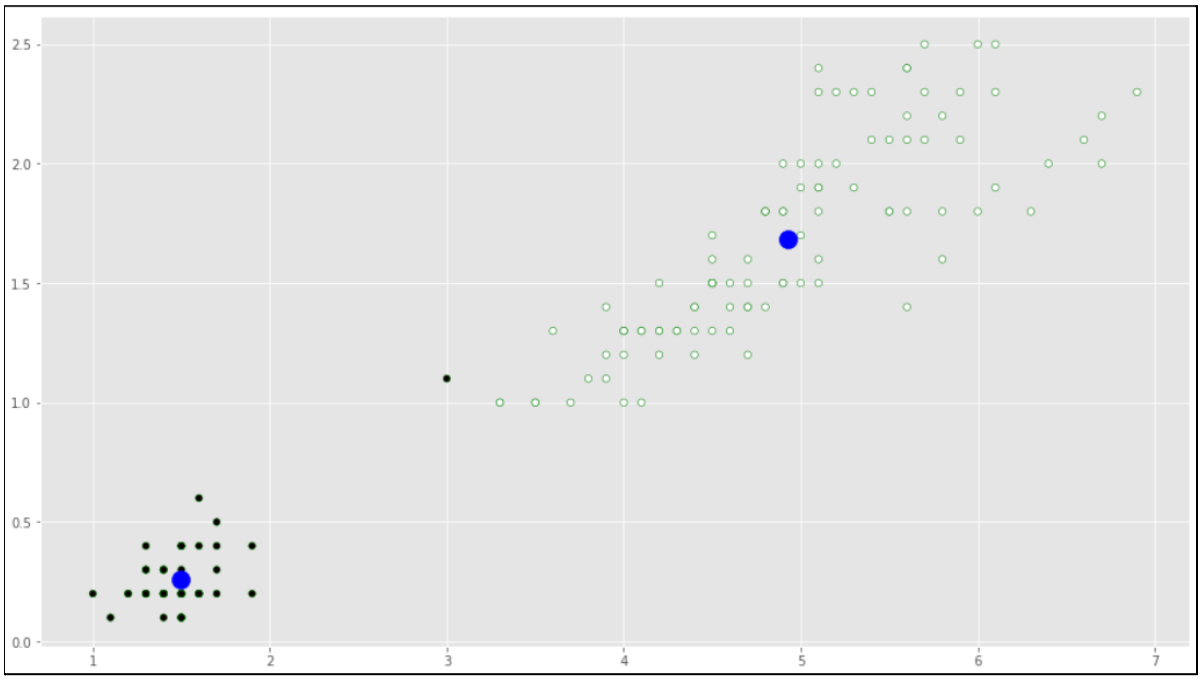
2.4 Clustering

Langkah berikutnya adalh melakukan *clustering* itu sendiri. *Clustering* akan dilakukan dengan menggunakan fungsi dari library sklearn, yaitu KMeans dengan jumlah klaster 2 (sepal *length* & sepal *width*, petal *length* & petal *width*), dan algoritma yang di atur ‘full’ untuk sepal serta ‘elkan’ untuk petal.

Berikut hasil dari *clustering* sepal:



Berikut hasil dari *clustering* petal:



2. 5 Evaluation

Langkah terakhir adalah melakukan evaluasi terhadap model yang telah dibuat. Evaluasi dilakukan dengan menggunakan *inertia* dan *silhouette coefficient*. Untuk *inertia*, akan dicari nilai dimana perubahan nilai 'k'-nya sudah tidak terlalu besar/signifikan. Sedangkan untuk silhouette, semakin hasilnya mendekati 1 maka semakin bagus dan semakin hasilnya mendekati -1 maka semakin jelek.

Hasil evaluasi sepal

```
+-- INERTIA --+
k:  1  COST:  130.18093333333334
k:  2  COST:  57.982406042078765
k:  3  COST:  37.12370212765957
k:  4  COST:  27.961759657351315
k:  5  COST:  21.077101654961503
k:  6  COST:  17.72402486772487
k:  7  COST:  14.723644130102675
k:  8  COST:  12.7037857975358
k:  9  COST:  11.16093971237741
```

Jika dilihat dari hasil tersebut, nilai 'k' sudah tidak terlalu besar perubahannya terjadi pada 'k' ke-4, maka jumlah klaster yang efektif adalah 4 klaster.

```
❄ Silhouetter score:  0.38177487325255094
```

Nilai *silhouette* agak sedikit mendekati 1 daripada -1, maka model ini lumayan bagus dalam melakukan *clustering*.

Hasil evaluasi petal

```
+-- INERTIA --+
⚡k:  1  COST:  550.6434666666667
⚡k:  2  COST:  86.40394533571003
⚡k:  3  COST:  31.38775897435898
⚡k:  4  COST:  19.48238901098901
⚡k:  5  COST:  13.933308757908758
⚡k:  6  COST:  11.056639971910453
⚡k:  7  COST:  9.213817958598739
⚡k:  8  COST:  7.688762403043182
⚡k:  9  COST:  6.496659206692712
```

Nilai ‘k’ yang perubahannya sudah mulai tidak signifikan ada pada k ke-4, maka jumlah klaster yang efektif adalah 4 klaster.

```
⚡Silhouetter score:  0.7651755502866581
```

Nilai *silhouette* mendekati 1 maka model ini sudah bagus dalam melakukan *clustering* petal.

BAB III PENUTUP

3.1. Kesimpulan

Pada praktikum pertemuan ketiga di mata kuliah pembelajaran mesin kali ini, mempelajari mengenai cara melakukan *clustering* untuk data yang tidak memiliki kelas atau label. Dataset yang digunakan pada praktikum ini adalah dataset bunga iris. Klaster yang digunakan dibagi menjadi dua, yaitu berdasarkan sepal dan berdasarkan petal. Untuk matrik evaluasi yang digunakan adalah *inertia* dan *silhouette coefficient*.

DAFTAR PUSTAKA

Hoseinzade, E., & Haratizadeh, S. (2019). CNNpred: CNN-based stock market prediction using a diverse set of variables. Expert Systems with Applications, 129, 273-285.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12, 2825–2830.*

Mitchell, Tom (1997). *Machine Learning*. New York: McGraw Hill. ISBN 0-07-042807-7. OCLC 36417892.

Hu, J.; Niu, H.; Carrasco, J.; Lennox, B.; Arvin, F., "Voronoi-Based Multi-Robot Autonomous Exploration in Unknown Environments via Deep Reinforcement Learning" IEEE Transactions on Vehicular Technology, 2020.

"What is Machine Learning?". *www.ibm.com*. Retrieved 2021-08-15.

Zhou, Victor (2019-12-20). "Machine Learning for Beginners: An Introduction to Neural Networks". *Medium*. Retrieved 2021-08-15.