T.C.
FATİH SULTAN MEHMET VAKIF ÜNİVERSİTESİ

Mühendislik Fakültesi
Bilgisayar Mühendisliği Bölümü

VERİ BİLİMİNE GİRİŞ DERSİ
FİNAL PROJESİ

HAZIRLAYAN:

Merve Özdemir
1421221003
Ümmügülsüm Can
1421221019

DANIŞMAN:
Dr. Öğr. Üyesi Ayla GÜLCÜ

1 - Proje Özeti

Bu veri seti, kentsel atık su arıtma tesisinde günlük sensör ölçümlerinden gelen verileri içermektedir. Veri seti çoklu labelların tahmin edilmesi işlemi için kullanılmaktadır. Fakat sınıflandırmadan farklı olarak labellar bilinmediği için Clustering uygulaması yapılacaktır

Amaç, arıtma işleminin her bir aşamasında tesisin durum değişkenleri ile arızaları tahmin etmek için tesisin işletim durumunu kümelemektedir. Bu etki alanı, yapılandırılmamış bir etki alanı olarak belirtilmiştir.

Analizin ana bulgusu ise suyun filtrelendikten sonraki kalitesinin farklı aralıklarda kümelenmesi işlemidir.

Veri seti: http://archive.ics.uci.edu/ml/datasets/Water+Treatment+Plant

2 - Veri Kümesi Açıklaması

- Veri kümesi örnek sayısı 527'dir.
- Veri kümesinde öznitelik sayısı 38'dir.
- Bu değişkenlerin her birinin türü object,int64,float (numeric, continuous)olarak değişkenlik göstermektedir.

3 - Temel Tanımlayıcı İstatistikler

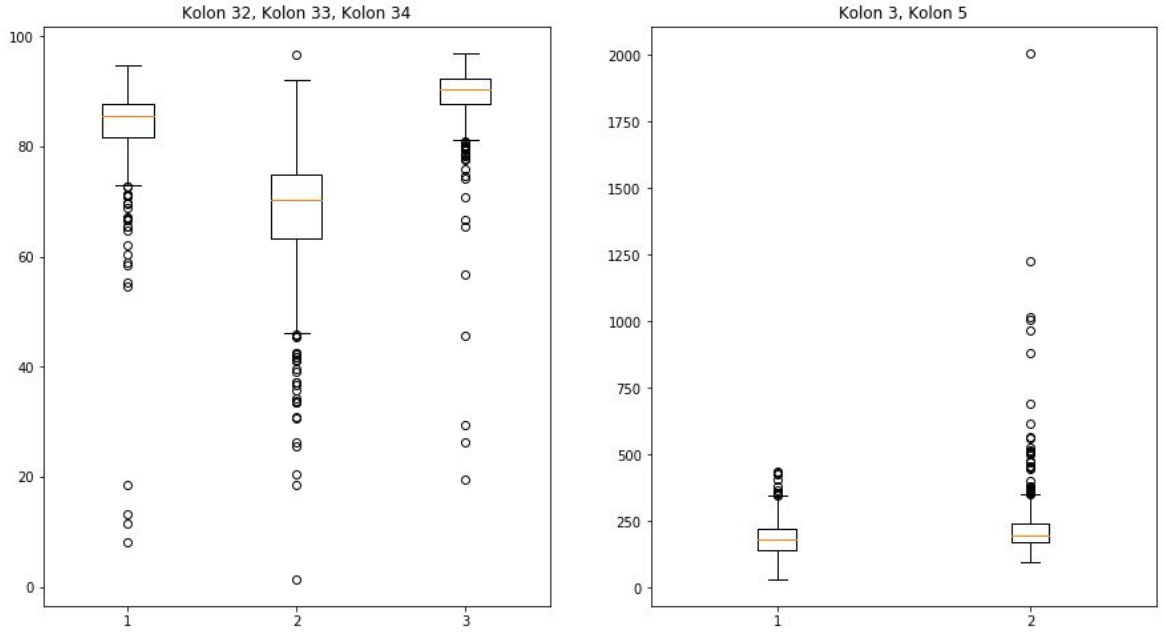
- Dağılım, minimum, maksimum,ortalama yüzdelik değerleri aşağıdaki tablodadır. Tablodan da görüleceği gibi max ve min değerlerinin ortalamadan sapma miktarları kolonlar içinde outlier değerleri olduğunu göstermektedir:

	0	1	2	3	4	5	6	7
count	527.000000	527.000000	527.000000	527.000000	527.000000	527.000000	527.000000	527.000000
mean	36960.886148	2.351328	7.810057	186.282732	406.592030	227.339658	61.503985	4.589374
std	6673.033177	2.743567	0.246175	61.750924	119.708608	135.832780	12.308208	2.678019
min	10050.000000	0.100000	6.900000	31.000000	81.000000	98.000000	13.200000	0.400000
25%	32557.500000	0.900000	7.600000	139.500000	326.500000	170.000000	55.850000	3.200000
50%	35729.000000	1.500000	7.800000	179.000000	397.000000	196.000000	64.500000	4.500000
75%	41094.000000	3.000000	8.000000	222.000000	474.500000	242.000000	69.600000	5.500000
max	60081.000000	33.500000	8.700000	438.000000	941.000000	2008.000000	85.000000	36.000000

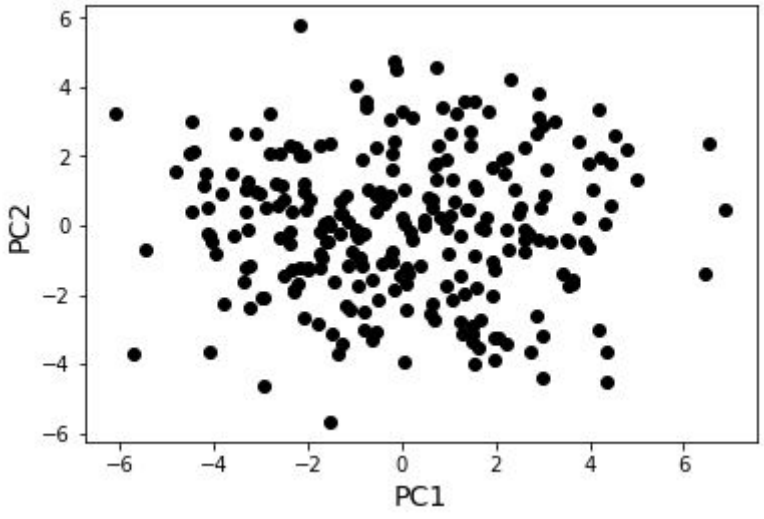
- Bir kaç kolonun eksik değerleriyle ilgili istatistikler % cinsinden tabloda görülmektedir:

Kolon 2	Kolon 3	Kolon 4	Kolon 5	Kolon 6	Kolon 7
3.415560	0.569	0.000	4.364	1.138	0.189

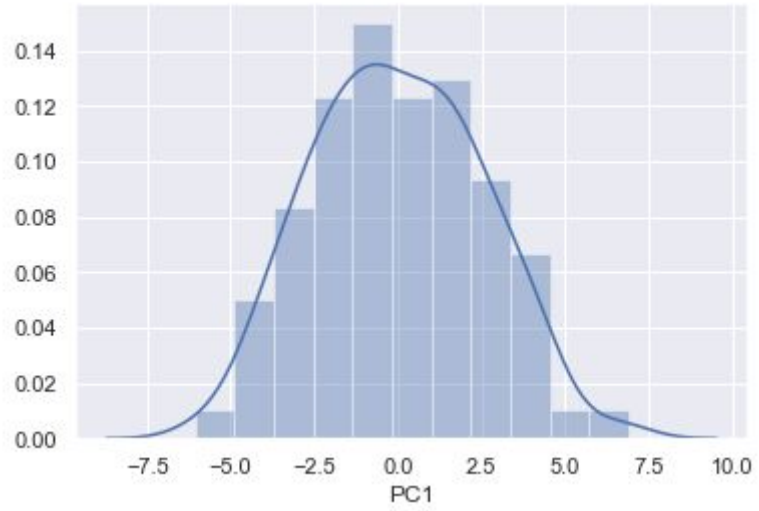
- Box-blot yöntemi ile aykırı değerler tespit edildi. Aşağıdaki box-plot çizimlerinde 5 kolonun bulundukları aralık ve outlier değerleri görülmektedir.

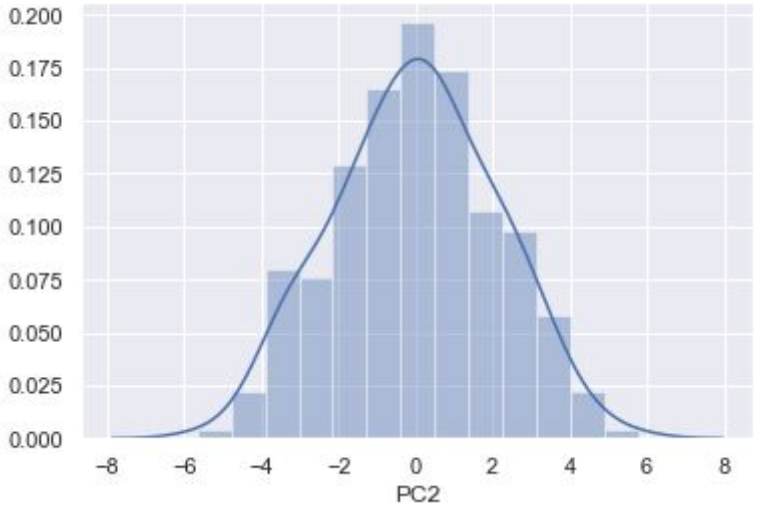


- Veri temizleme işlemleri gerçekleştirildikten sonra 2 boyuta indirgenmiş ve verinin dağılım grafiği aşağıdaki şekildedir:



- İndirgenen kolonların dağılımları normal dağılıma sahip oldukları görülmektedir:





4- Veri Yönetimi Süreçleri

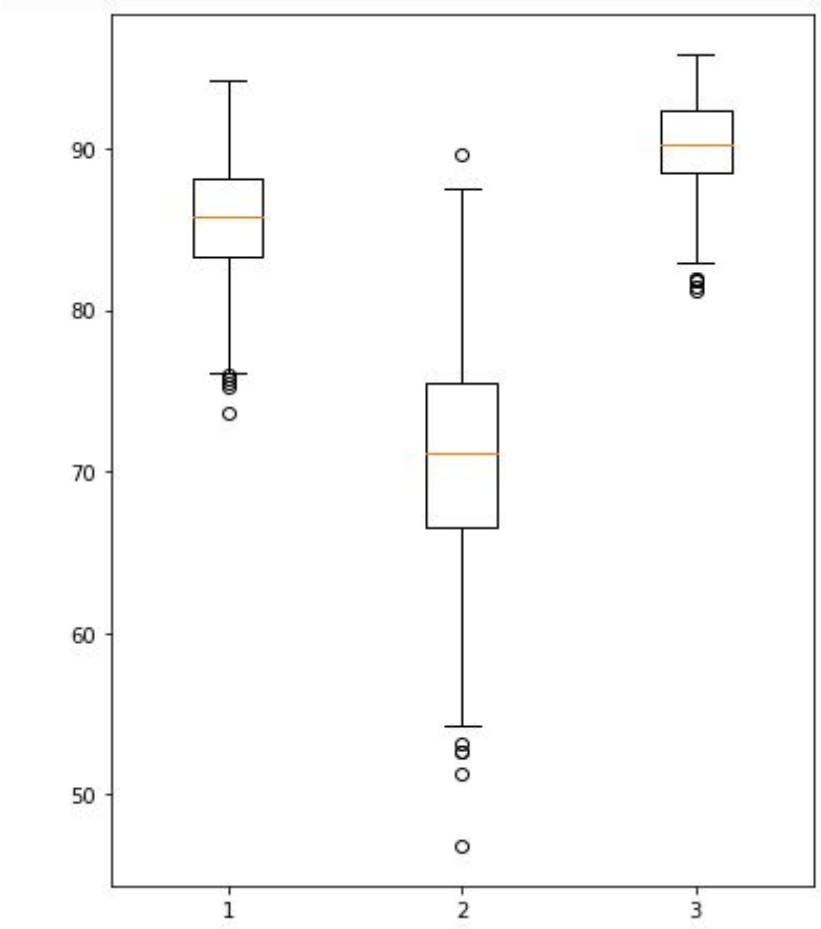
- Dataset üzerinde az etkiye sahip olan ilk kolon (Tarih) silinmiştir.
- Missing değerlerin temizlenme işlemi aşamasında, NaN değerler yerine her kolonun kendi mod (en çok tekrar eden) değerleri konulmuştur.
- Öznitelik türleri float ve int64 türündedir fakat missing değerler sebebiyle object olan kolonlar vardır. Bu kolonlar missing değerlerin düzeltilmesi işleminden sonra continuous değerlere dönüştürülmüştür.

- Outlier Detection işlemi Box-plot aracılığıyla yapılmıştır. Bu değerlerin veri setinden çıkarılmasında IQR yöntemi kullanılmıştır. Verinin %25'i Q1, % 75'i Q3 olarak adlandırılır. Bu iki değer arasındaki fark bize IQR değerini verir. Alt uç değer ve üst uç değer hesabının 1.5 kat dışında kalan değerler outlier olarak alınır. Data setinden bu değerler çıkarılmıştır

IQR = Q3 - Q1

alt sınır : Q1 - 1.5 * IQR

üst sınır: Q3 + 1.5 * IQR

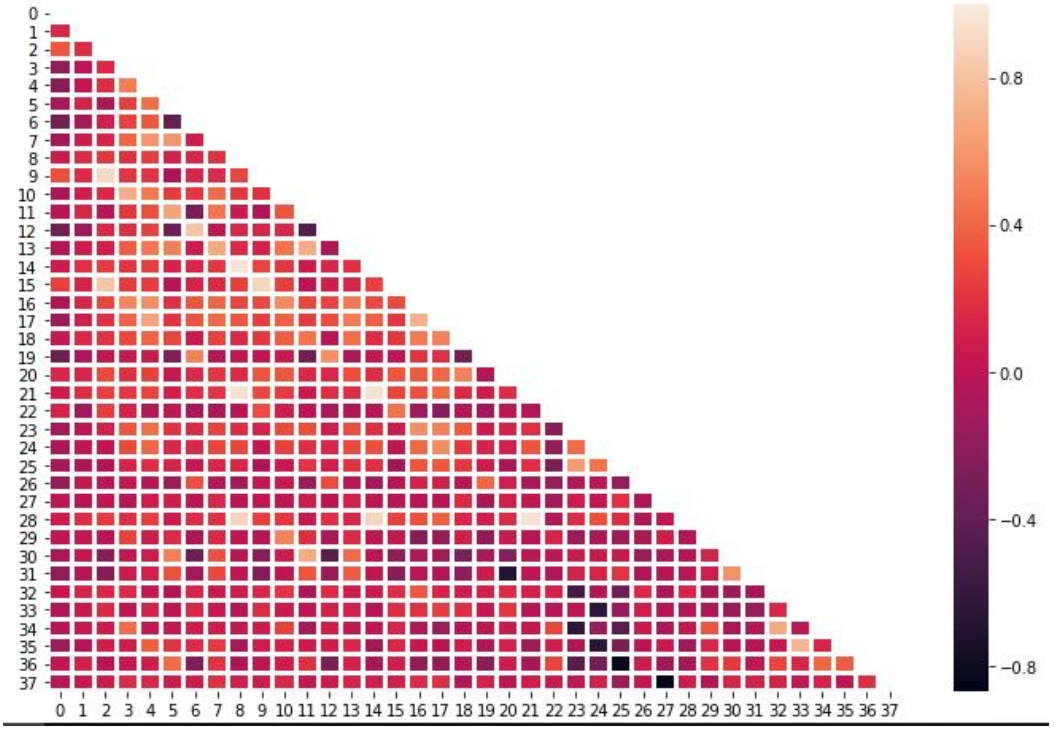


Yukarıdaki box-plot outlier değerler atıldıktan sonraki halidir.

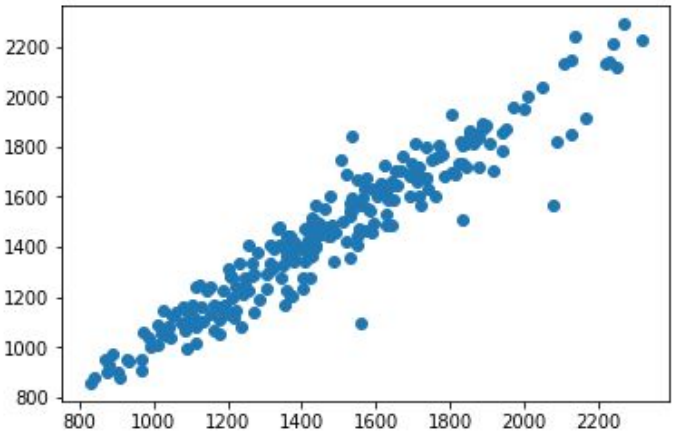
• Feature Selection işlemi için data setinin korelasyon matrisine bakılmıştır ve birbirleriyle ilişki olan değerler manuel olarak atılmıştır. Daha sonra Feature Extraction yaparak datayı hem normalize (scaling) edip hem de data üzerinde boyut azaltmak için kullanılmıştır. Bu işlem PCA yöntemi ile yapılmıştır.

5-Veri görüntüleme

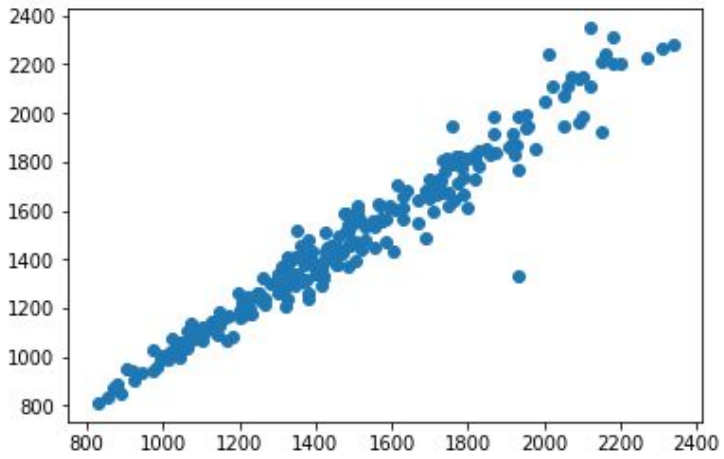
Değişkenlerin her bir çifti arasındaki ilişki korelasyon matrisinde görselleştirilmiştir.



- Veri setindeki 21. ve 28. kolonların pozitif ilişkide oldukları görülmektedir.



- Veri setindeki 14. ve 8. kolonların pozitif ilişkide oldukları görülmektedir.



Bu grafiklerden yola çıkarak feature selection el ile yapılmıştır.

6- Makine Öğrenmesi (ML) Uygulaması:

Model Seçimi ve Değerlendirme

Bu proje için Clustering yöntemi kullanılmıştır. Clustering label olmadan belirlenen bazı algoritmalarla gruplama işlemi yapılmaktadır.
Clustering uygulaması için **K-Means ve Hiyerarşik Clustering** yaygın olarak kullanılan kümeleme algoritmalarıdır. Çalışmamızda bu algoritmalar kullanılmıştır.

1. K-MEANS:

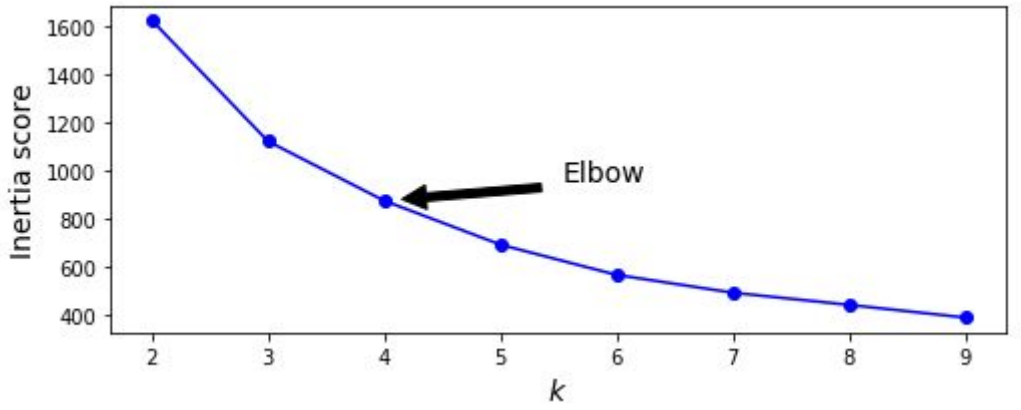
Bu algoritma için kullandığımız metrikler:interia ve silhoutte score.Bu iki metrik kümeleme yönteminin ne kadar iyi grupladığını analiz etmek için iki farklı ölçümde bakmamızı sağladı.

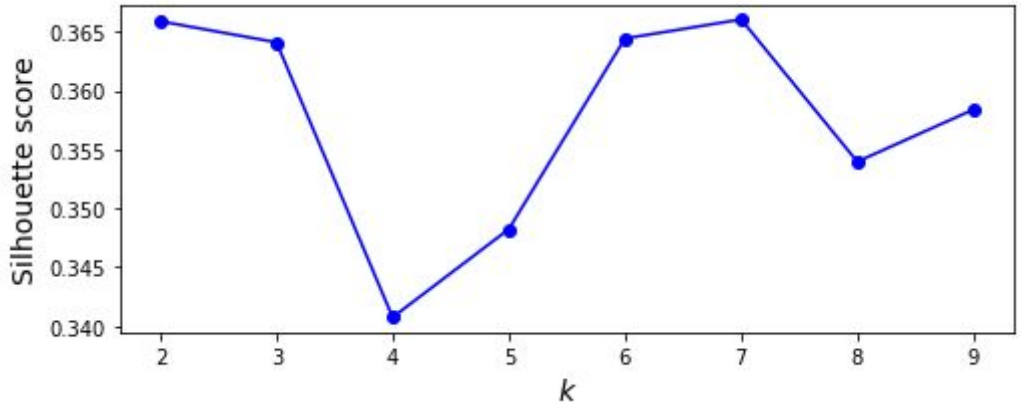
Inertia; Her küme için karesel hatanın toplamıdır. Bu nedenle atalet ne kadar küçük olursa küme o kadar yoğunlaşır.(tüm noktalar birbirine yakınsa)

Silhoutte score: -1'den 1'e kadardır ve kümelerin birbirinden ne kadar uzak veya uzak olduğunu ve kümelerin ne kadar yoğun olduğunu gösterir. Siluet puanınız 1’e ne kadar yakınsa kümeleriniz o kadar belirgindir.

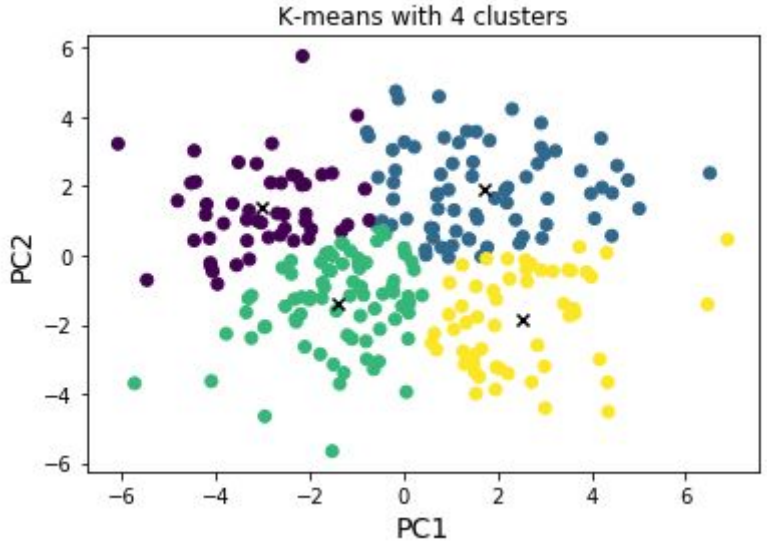
$$S = (b - a) / \max(a, b)$$

Alınan sonuçlar:
Aşağıdaki interia ve silhoutte grafiği en iyi k değerini belirlemekte yardımcı oluyor. k = 4'de en iyi kümeleme işlemi gerçekleştirilmiştir.

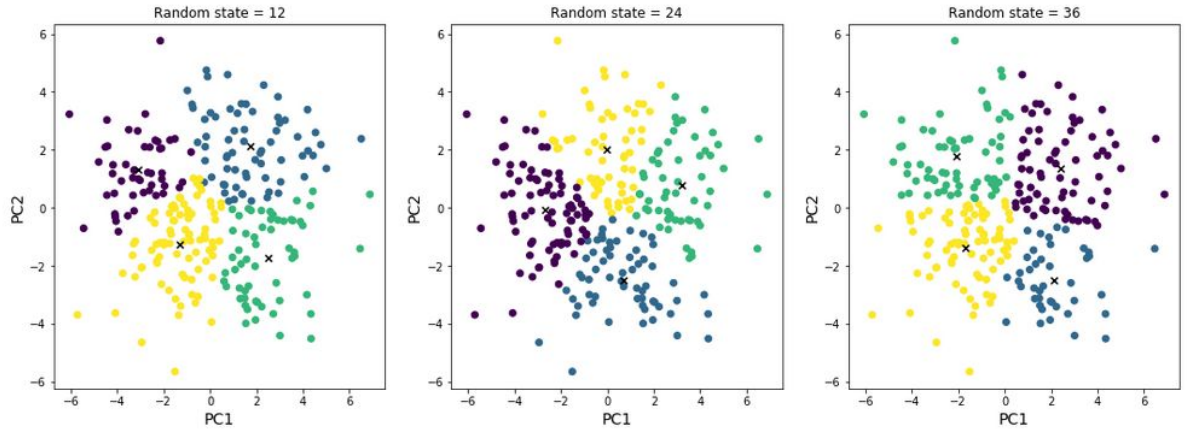




Belirlenen k değeriyle yapılmış kümele işleminin grafiği aşağıda verilmektedir:



Aşağıda ise farklı random state değerleri ile oluşturulan kümeler verilmektedir:

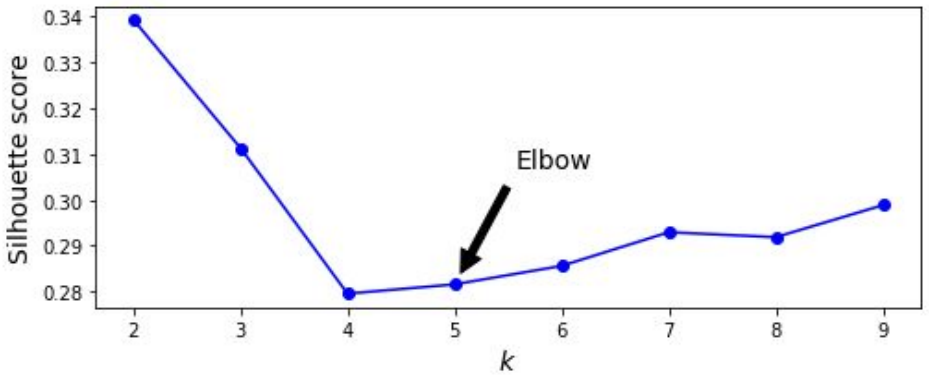


2. YIĞINSAL (AGGLOMERATIVE) KÜMELEME (HİYERARŞİK KÜMEMELE):

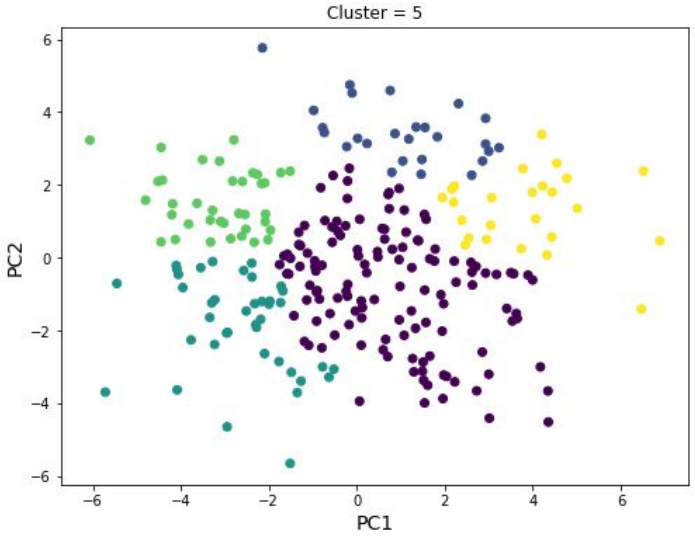
Hiyerarşik kümelemede yukarıdan aşağıya doğru önceden belirlenmiş sıraya sahip kümeler oluşturulur. Agglomerative Clustering de bunun bir türüdür.

Alınan sonuçlar:

Aşağıdaki interia ve silhoutte grafiği en iyi k değerini belirlemekte yardımcı oluyor. k = 5'de en iyi kümeleme işlemi gerçekleştirilmiştir.

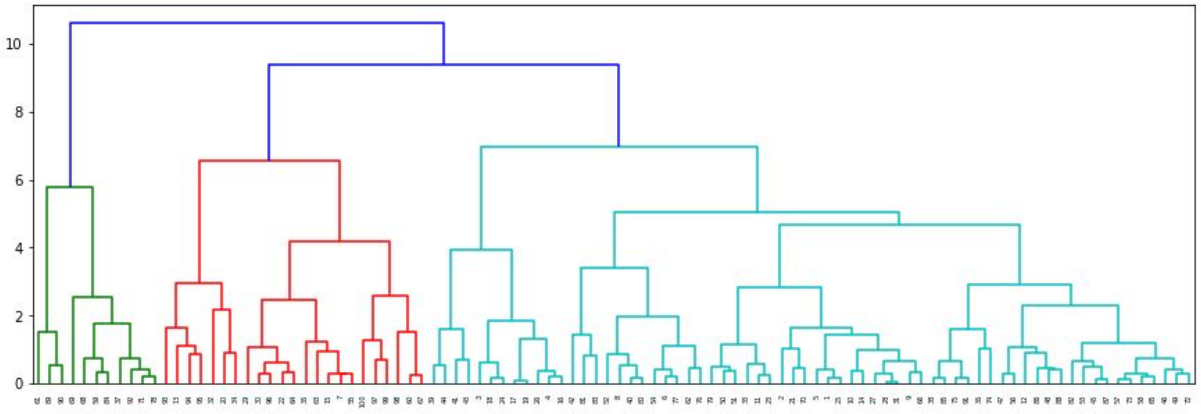


Belirlenen k değeriyle yapılmış kümele işleminin grafiği aşağıda verilmektedir.

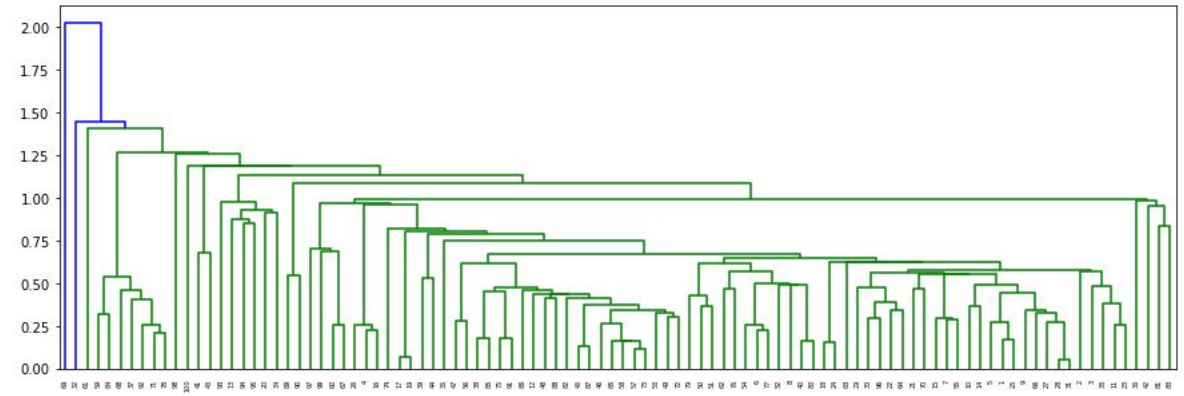


Veri setindeki cluster değerini belirlemek için dendogram grafiği kulanılmıştır.

Linkage =complete

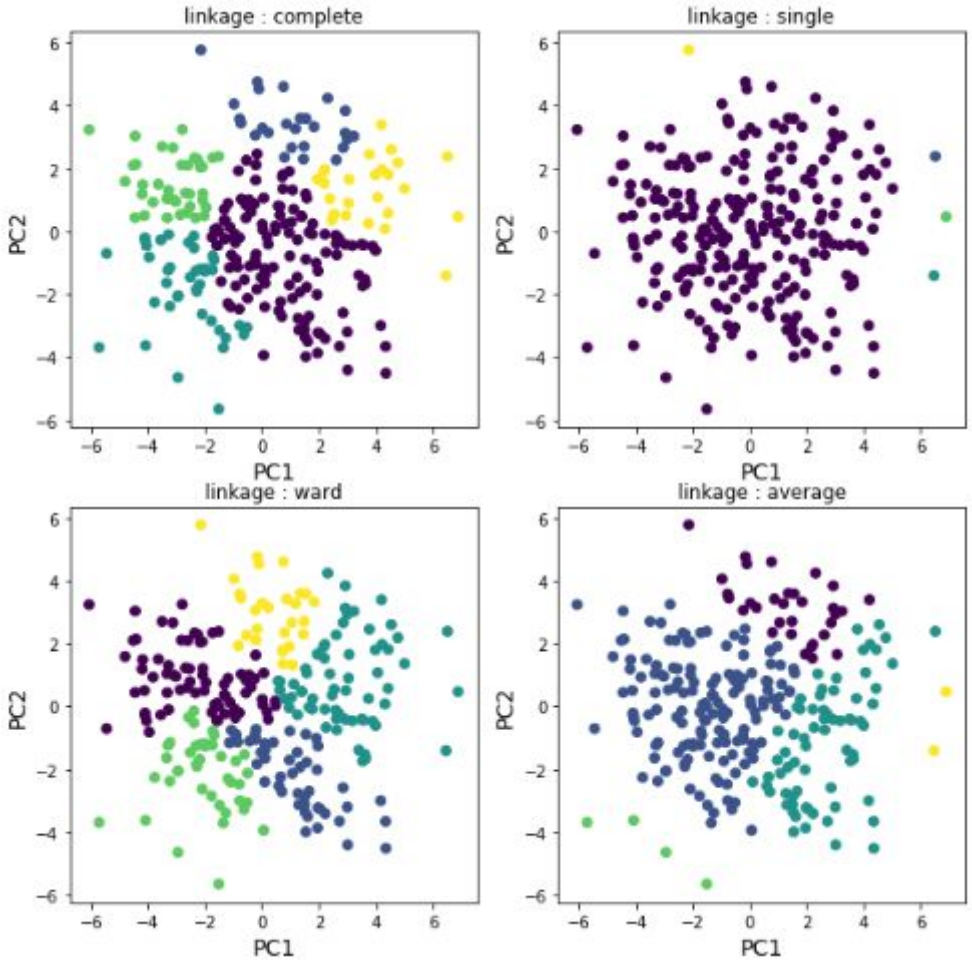


Linkage = single



Görüldüğü gibi complete link single'a göre daha iyi sonuç vermiştir.

Aşağıda ise Agglomerative algoritması kullanılarak farklı linkage değerleri ile oluşturulan kümelerin grafikleri verilmektedir (k=5):



Bu grafiklerden yola çıkarak en iyi linkange parametresinin ward ve complete olduğu görülmektedir.

SONUÇ

Bulunan en iyi parametrelerle birlikte (KMeans için cluster =4 ve Agglomerative için cluster = 5, linkage = ward) alınan silhoutte score değerlerine bakarak KMeans algoritmasının daha iyi geldiği görülmektedir.

```
print('silhoutte score: ', silhouette_score(data_2d, kmeans.labels_))
```

```
silhoutte score:  0.3407160264839394
```

```
print('silhoutte score: ', silhouette_score(data_2d, agg_2.labels_))
```

```
silhoutte score:  0.2815305984514455
```