PATIENT CLUSTERING

I.GİRİŞ

Veri kümesi acil servise başvuran hastaların tıbbi verilerini içermektedir. Hasta semptomlarına dayanarak, acil müdahaleye ihtiyaç duyan hastaların belirlenip; hastaların önceden belirlenmiş bir hasta bakım alanına yönlendirilerek bakımlarına öncelik verilip uygun şekilde tedavi önlemlerinin başlatılmasına olanak tanıyacak modeller geliştirilmesi istenmektedir.

Veri Setinin Temel Özellikleri:

Veri kümesi 6000 hastadan alınan sonuçlardan oluşmaktadır. 6000 satır ve 16 sütun bulunmaktadır. Veri seti nümerik ve kategorik değişkenlerden oluşuyor. Ayrıca bazı eksik değerler bulunmakta. Hangi ölçümlerin kullanıldığı aşağıdaki tabloda verilmiştir:

age – Yaş

gender – Cinsiyet

chest_pain_type – Göğüs Ağrısı Türü

blood_pressure – Kan Basıncı

cholesterol – Kolesterol

max_heart_rate – Maksimum Kalp Atış Hızı

exercise_angina – Egzersiz Anjinası

plasma_glucose – Plazma Glukoz Seviyesi

skin_thickness – Deri Kalınlığı

insulin – İnsülin Seviyesi

bmi – Vücut Kitle İndeksi

diabetes_pedigree – Diyabet Aile Geçmişi Skoru

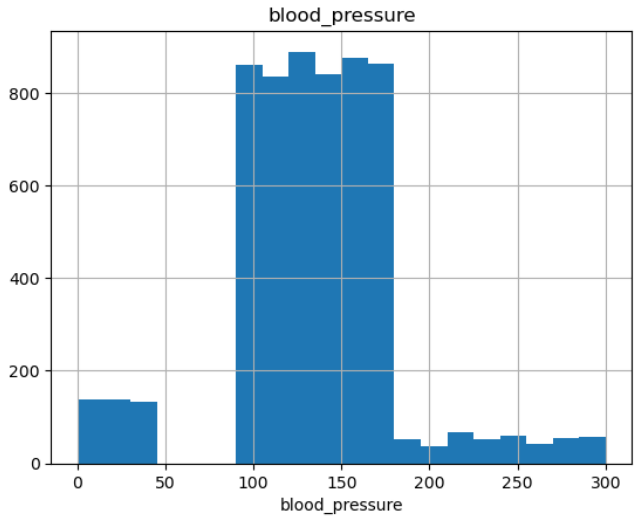
hypertension – Hipertansiyon

heart_disease – Kalp Hastalığı

residence_type – Yaşadığı Bölge Türü

smoking_status – Sigara Kullanım Durumu

Kodda değişkenlerin veri seti üzerindeki dağılımları şekildeki gibi görselleştirilmiştir ve bunların ortalama, standart sapma gibi istatistiksel özellikleri açıklanmıştır.




```
count   6000.000
mean     209.845
std       52.840
min      120.000
```

Sınıf (kategori) dağılımını anlamak, sık görülen kategorileri tespit etmek ve eksik veya nadir görülen sınıfları belirlemek için kategorik ve nümerik değişkenlerin analizini yaptım.

```
                smoking_status  Ratio
smoking_status
Smoker                    2786 46.433
Non-Smoker                2738 45.633
Unknown                    476  7.933
        gender  Ratio
gender
0.000     2777 46.283
1.000     2751 45.850
```

II.GELİŞME

VERİ ÖNİŞLEME

Veri setinde önce eksik verilerin sayısını ve oranlarını tespit ettik daha sonra eksik veriler tüm verilere oranla az sayıda olduğu düşünerek bu verilerin olduğu satırları kaldırdım. Bu sayede herhangi bir eksik değerimiz kalmadı.

Kategorik verileri (residence_type ve smoking_status) label encoding yöntemiyle sayısal değerlere yani 0 ve 1 değerlerine dönüştürdüm.

Ki-Kare Bağımsızlık Testi kullanarak residence_type ve smoking_status sütunlarının diğer değişkenlerle istatistiksel bağımsız olup olmadığını analiz ettim.

```
Çapraz Tablo (hypertension):
hypertension     0   1
residence_type
0              961  900
1              932  941
Chi2 Sonucu (attribute hypertension): Chi2=1.2446986195146 49, p-value=0.26456743303791624, dof=1
```

Örneğin yukarıdaki çıktıya göre;

- residence_type ve hypertension bağımsızdır.
- p-değeri 0.2646 olduğundan, H0 hipotezini reddedemiyoruz. Bu, residence_type (ikamet türü) ve hypertension (hipertansiyon) arasında istatistiksel olarak anlamlı bir bağımlılık olmadığını gösterir.
- 1.2447, çok düşük bir değerdir. Bu, gözlenen ve beklenen frekanslar arasında büyük bir fark olmadığını destekler.

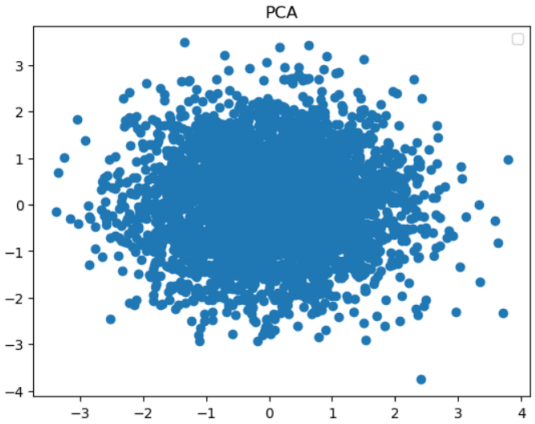
residence_type değişeni diğer niteliklerle zayıf veya hiç anlamlı ilişki göstermediğinden, sütun kaldırılacaktır.

Veri kümesindeki kategorik sütunları kodlayarak ve ardından tüm sütunları standartlaştırarak, veriyi makine öğrenimi algoritmalarına uygun hale getirdik. Kategorik sütunların sayısal değerlere dönüştürülmesi ve tüm sütunlardaki verilerin ortalama 0 ve standart sapma 1 olacak şekilde dönüştürülmesi (z-skor normalizasyonu) amaçlanmıştır.

Daha sonra yine kategorik sütunların sayısal değerlere dönüştürülmesi ve verilerin belirli bir aralıkta yeniden ölçeklendirilmesi (bu örnekte [-1, 1] aralığı) için min-max scaler kullandık.

K MEANS İLE KÜMELEME

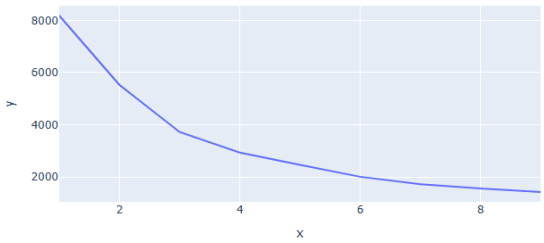
İlk başta PCA ile verinin boyutunu iki ana bileşene indirgedik.

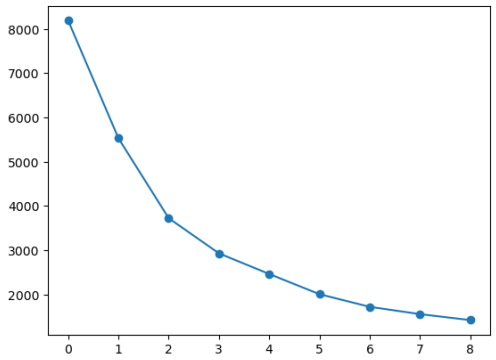


Bu grafik veri yapısının genelde simetrik bir dağılım göstermekte ve çok yoğun olduğunu gösteriyor.

Daha sonra K-Means algoritması ile optimum küme sayısı belirlemeye çalıştık. K-Means algoritması, kümeleme için kullanılır ve bu durumda veri setindeki hastalar, tıbbi semptomlarına ve aciliyet durumlarına göre gruplandırılabilir.

Dirsek yöntemi ise kümelerin sayısını belirler. Elde edilen grafikte "dirsek" noktası, ideal küme sayısını gösterir. Bu, hastaların uygun şekilde kategorize edilmesini kolaylaştırır.





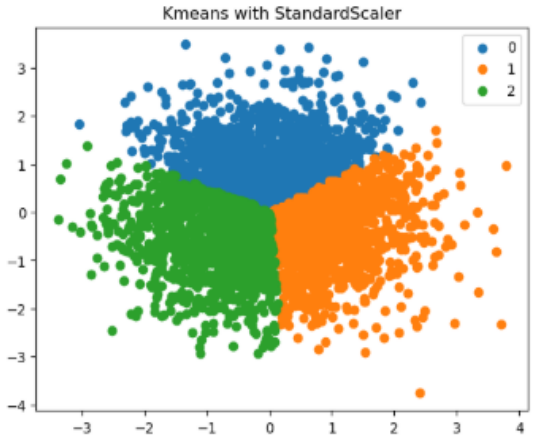
Elde ettiğiniz grafikteki WCSS eğrisine bakıldığında: Küme sayısının 3 civarında optimal olduğu görülüyorsa, bu durum şunu ifade eder:

Küme 1: Acil müdahale gerektiren kritik hastalar.

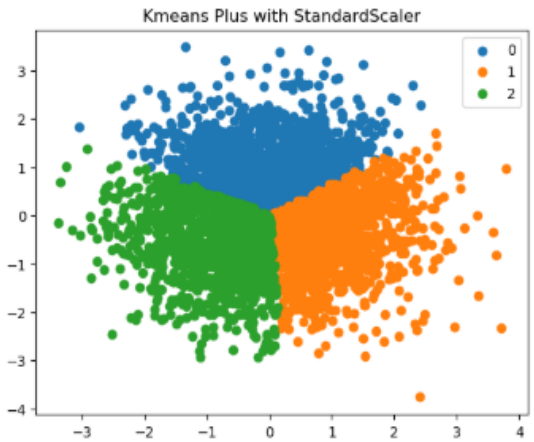
Küme 2: Acil müdahale gerektirmeyen, ancak hızlı değerlendirme gerektiren hastalar.

Küme 3: Durumu stabil olan ve acil müdahale gerekmeyen hastalar.

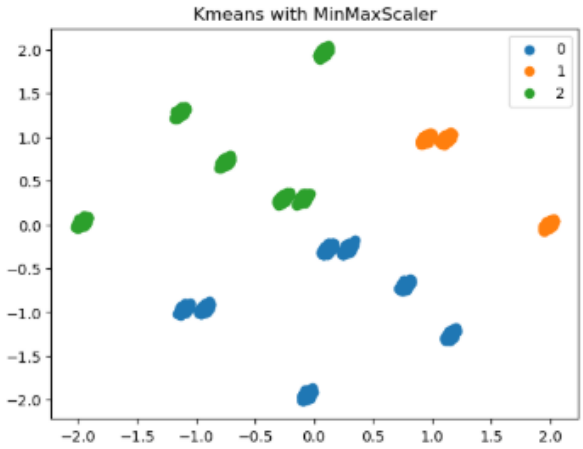
Hastaların verilerini k-means ve k-means++ algoritmaları kullanılarak kümeleme sonuçlarını inceledik. Inertia (Kümeleme İçindeki Toplam Sapma, Silhouette Score (Silüet Skoru), Davies-Bouldin Score metrikleriyle sonuçları inceledik.



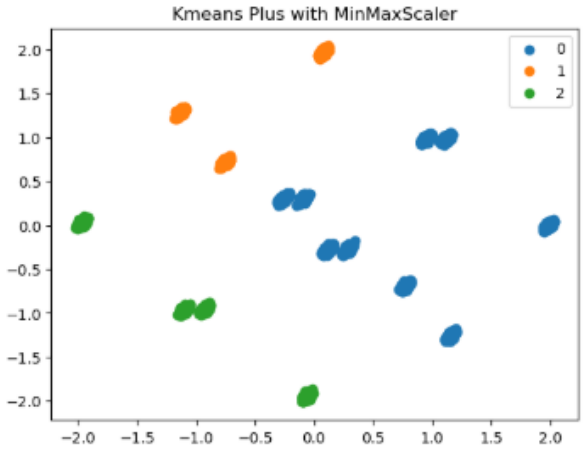

```
Inertia k-Means with StandardScaler: 3769.1179570358036
Silhouette_score k-Means with StandardScaler : 0.32381885729941473
Davies_bouldin_score k-Means with StandardScaler : 0.9473479924202546
```




```
Inertia k-Means++ with StandardScaler: 3683.896971301311
Silhouette_score k-Means++ with StandardScaler: 0.3348144309663329
Davies_bouldin_score k-Means++ with StandardScaler : 0.929871953866985
```




```
Inertia k-Means with MinMaxScaler: 3134.9940309700405
Silhouette_score k-Means with MinMaxScaler : 0.4347855265872827
Davies_bouldin_score k-Means with MinMaxScaler : 0.872643277027814
```




```
Inertia k-Means++ with MinMaxScaler: 3272.6874301662486
Silhouette_score k-Means++ with MinMaxScaler: 0.4232467122717317
Davies_bouldin_score k-Means with MinMaxScaler : 0.872643277027814
```

StandardScaler, veriyi ortalaması 0 ve varyansı 1 olacak şekilde standardize eder. Sonuçlara göre:

- K-Means++ algoritması daha tutarlı bir performans sergilemiş.
- Silhouette skorları kümeler arasında ortalama bir ayrım olduğunu gösteriyor.

MinMaxScaler, veriyi 0 ve 1 arasında yeniden ölçeklendirir. Sonuçlar, farklı ölçekleme yöntemlerinin kümeleme sonuçlarını etkileyebileceğini göstermektedir. Özellikle K-Means++ burada da avantajlı görünmekte. Ancak Silhouette skorları, kümeleme ayrımında daha düşük bir performans olduğunu göstermiş.

K-Means++ algoritması, hem StandardScaler hem de MinMaxScaler ile daha optimize sonuçlar vermiş. Bu nedenle, genelde K-Means++ önerilir. StandardScaler ile elde

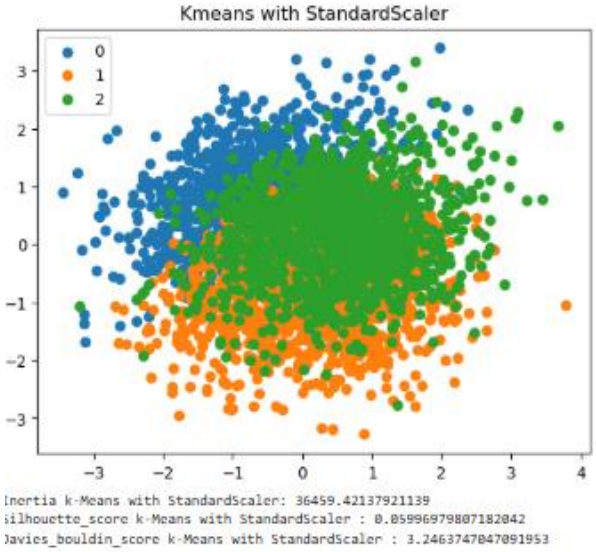
edilen sonuçlar daha net küme ayrımı göstermekte.

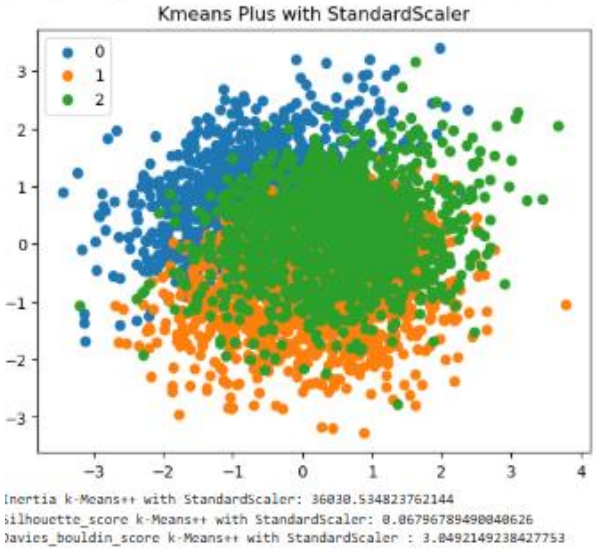
Daha sonra optimize edilmiş PCA'nın etkisini görmek için K Means'i optimize ettik. gender, smoking_status, bmi, skin_thickness sütunları çıkarılmış ve manuel olarak belirlenen önemli sütunlar seçilmiş (heart_disease, hypertension, diabetes_pedigree, vb.). Bu özellik seçimi, veri kümesinin boyutunu düşürmüştür. Daha "anlamlı" ve "temiz" bir veri kümesi elde edilir. Veri kümesinin boyutu indirgendikten sonra yoğunluğunda bir azalma görüldü.

Bizim problemimizde bu sayede azalan yoğunluk, modelin gerçekten önemli olan tıbbi faktörlere (ör. yaş, tansiyon, kalp hastalığı) odaklanmasını sağlar. Modelin daha hızlı çalışmasını ve tıbbi karar süreçlerini destekler. Model, az sayıda fakat anlamlı özellikler kullanarak farklı hasta gruplarında daha iyi genelleme yapabilir. Ancak, yoğunluk azalırken eğer gereğinden fazla bilgi çıkarılırsa, model için önemli olabilecek bazı ilişkiler kaybolabilir.

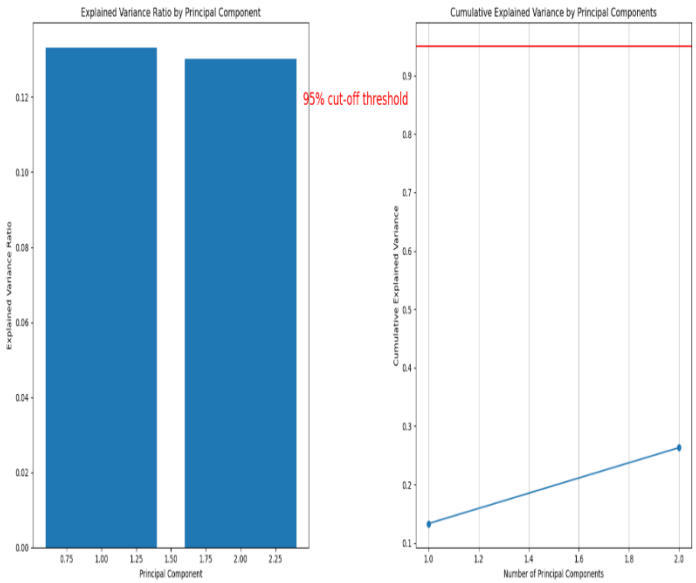
Optimize edilmiş verilerle de "elbow yöntemi" kullanılarak optimal küme sayısının belirledik. Bu analiz sonucunda da, küme sayısını 3 olarak seçmek anlamlı görünmektedir.

- K-Means ve K-Means++ algoritmaları kullanılarak, StandardScaler ile ölçeklendirilmiş bir veri kümesinde 3 küme oluşturulmuştur.
- Silhouette skorunun düşük olması, veri kümelerinin net olarak ayrılmadığını ve sınıfların birbiriyle örtüştüğünü gösteriyor.
- Davies-Bouldin skorunun büyük olması da kümelerin birbirine yakın olduğunu ifade ediyor.
- K-Means++ başlangıç merkezlerini daha iyi seçtiği için, sonuçlar K-Means'e göre daha iyi.
- Inertia, Silhouette ve Davies-Bouldin skorlarında küçük ama anlamlı iyileşmeler var.
- K-Means++ rastgele başlangıca göre daha iyi performans göstermiştir. Ancak sonuçlar halen düşük kaliteli kümeler oluşturmaktadır.
- Daha iyi performans için özellikler arasında daha anlamlı bir seçim yapılmalı veya boyut indirgeme teknikleri (PCA) uygulanmalıdır.





PCA analizi ile veri setindeki boyutları azaltarak verideki en önemli değişkenleri tespit etmeye çalıştık. Ancak, mevcut PCA analizi varyansın büyük bir kısmını açıklayamıyor (kümülatif varyans düşük). Bu da, daha fazla bileşene ihtiyaç duyulduğunu gösterir. Daha fazla bileşen eklendiğinde varyansın daha büyük bir kısmı açıklanabilir. Grafikte, 13 bileşenin %95'lik varyans eşiğini karşıladığı belirtilmiş. Bu nedenle, 13 bileşenli yeni PCA analizi bu durumu iyileştirebilir.

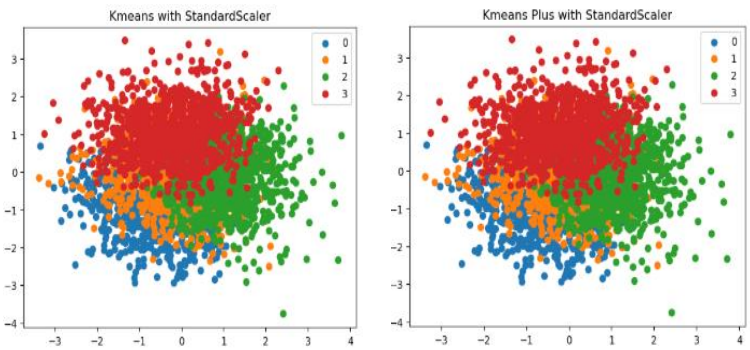


Bir defa daha dirsek yöntemi uyguladık ve bu sefer hastaları daha spesifik olarak farklı risk gruplarına ayrılabilmek için K=4 seçtik.

Daha sonra tekrar K-means ve K-means++ algoritmalarını kullanarak verileri kümeledik.

K-means++, klasik K-means'e göre biraz daha iyi ayrışma sağlıyor (Silhouette Score biraz daha yüksek, Davies-Bouldin Score daha düşük).

HİYERARŞİK YAKLAŞIM



Agglomerative Clustering (Hiyerarşik Kümeleme) kullanılarak farklı bağlantı (linkage) yöntemleri ve ölçekleme teknikleri ile sonuçları karşılaştırdık.

- **Single Linkage:** Küme içindeki en yakın noktalar arasındaki mesafeye odaklanır. Zincirleme bağlanma eğilimi gösterir.
- **Average Linkage:** Kümedeki tüm noktalar arasındaki ortalama mesafeyi baz alır.

Ölçekleme Yöntemleri

- **StandardScaler:** Veriyi ortalaması **0**, standart sapması **1** olacak şekilde standartlaştırır.
- **MinMaxScaler:** Veriyi **[0,1]** aralığına ölçekler.

Değerlendirme Metrikleri

- **Silhouette Score:** Küme içi tutarlılığı ve kümeler arası ayrışmayı ölçer. **1'e yakın değerler iyi kümelenmeyi gösterir**.
- **Davies-Bouldin Score:** Küme içi sıkılığı ve kümeler arası ayrımı değerlendirir. Düşük değerler daha iyi kümelenmeyi ifade eder.

StandardScaler ile Kümeleme Sonuçları:

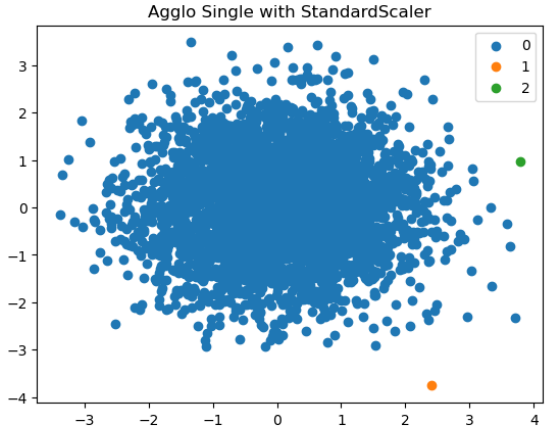
- *Single Linkage* yöntemi, büyük bir küme oluşturmuş ve birkaç noktayı dışarıda bırakmış. Bu, yönteminin kümeleme için pek etkili olmadığını gösteriyor.
- *Average Linkage* yöntemi, daha dengeli bir kümeleme yapmış gibi görünüyor, ancak kümeler yine net ayrılmamış.

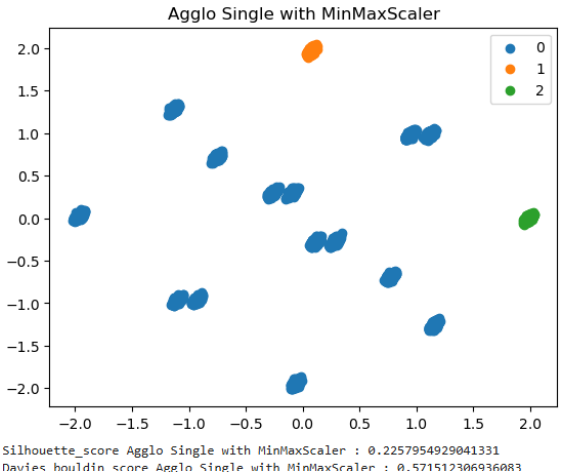
MinMaxScaler ile Kümeleme Sonuçları:

- Single Linkage yöntemiyle, verilerin az sayıda kümeye ayrıldığı görülüyor.

Kümeler, dağınık halde ve büyük olasılıkla kötü sonuç veriyor.

- Average Linkage yöntemi, daha belirgin kümeler oluşturmuş gibi görünüyor.





Single Linkage yöntemiyle Silhouette Score yüksek, ancak Davies-Bouldin Score oldukça düşük.

- Bu, kümelerin birbirine **oldukça yakın** olduğunu gösterir, ancak iç tutarlılığı yüksektir.
- Single Linkage **zincirleme bağlanma eğiliminde olduğundan**, büyük bir küme oluşturmuş olabilir.

Average Linkage ile Silhouette Score düşmüş, Davies-Bouldin Score ise yükselmiş.

- Bu, kümelerin daha dağınık olduğunu ve birbirinden net ayrılmadığını gösterir.
- Kümeler birbirinden yeterince ayrılmadığı için acil hastaların doğru belirlenmesi zor olabilir.

Single Linkage, Silhouette Score açısından en kötü sonucu verdi (0.2257).

- Verinin ölçeklenme biçimi, Single Linkage'ın zincirleme bağlanma eğilimini artırmış olabilir.
- Küçük kümeler yerine büyük ve birleşik kümeler oluşmuş olabilir.

Average Linkage, MinMaxScaler ile daha iyi bir Silhouette Score (0.4239) verdi, ancak Davies-Bouldin Score yüksek (0.9132).

- Bu, kümelerin iç yapısının daha uyumlu olduğunu ancak kümelerin birbirinden net ayrışmadığını gösteriyor.

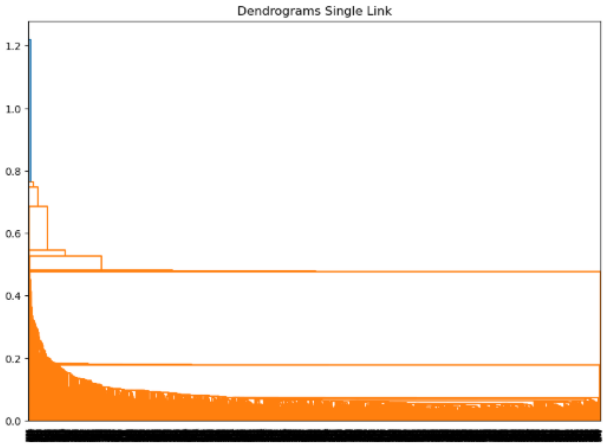
Modelin Acil Hasta Önceliklendirmesi İçin Yeterliliği

- Single Linkage kümeleri genellikle tek bir büyük küme oluşturduğu için, acil hastaları belirlemekte başarısız olabilir.
- Average Linkage biraz daha iyi çalışsa da, sonuçlar yeterince iyi değil.
- Silhouette Score'ların genel olarak düşük olması, kümeler arası ayrışmanın yeterli olmadığını gösteriyor.
- Davies-Bouldin Score'un bazı durumlarda yüksek olması, kümelerin birbiriyle fazla iç içe geçtiğini ve belirgin bir ayrım olmadığını gösteriyor.

Verinin küme yapısını anlamaya, kümeleme yöntemlerinin sonuçlarını görselleştirmeye, optimum küme sayısını belirlemeye ve hasta gruplarını anlamlı bir şekilde ayırmaya olanak tanıması için dendrogram kullandık. Bu sayede hangi hasta gruplarının benzer semptomlara veya tıbbi verilere sahip olduğu, kritik öneme sahip hasta gruplarının kaç farklı gruba ayrılması gerektiği, yüksek seviyelerde

geniş hasta grupları (örneğin "acil" ve "acil olmayan") gibi sonuçlara ulaşabiliriz.

Aşağıda örnek olması açısından iki farklı dendrogram verilmiştir.

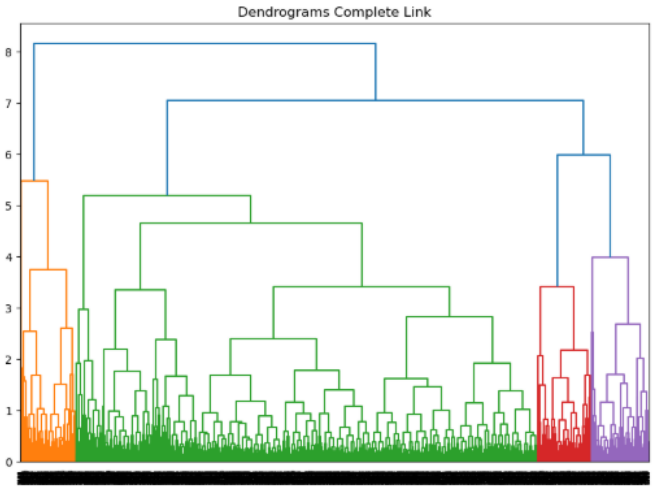


Single Linkage (En Yakın Komşu Yöntemi), iki küme arasındaki mesafeyi bu kümelerdeki en yakın noktalar arasında ölçer. Yukarıdaki dendrogram, bu yöntemin sonucudur ve kümeler arasındaki birleşme sırasını gösterir.

Grafiğin alt kısmı oldukça yoğun; bu, veri setindeki birçok veri noktasının birbirine çok yakın mesafelerde bulunduğunu gösterir.

Dikey çizgiler arasındaki büyük boşluklar, kümeler arasındaki mesafenin önemli derecede arttığını ifade eder. Özellikle üst bölümlerde büyük mesafeli birleşmeler gözlenir. Bu, veri setinin açıkça ayrılmış birkaç büyük kümeden oluşabileceğini işaret edebilir.

Single Linkage yöntemiyle yapılan bu analiz, belirli hasta gruplarının birbirine ne kadar yakın olduğunu ortaya çıkarabilir. Ancak daha dengeli bir kümeleme için Complete Linkage gibi diğer yöntemler de kullanılmalıdır.



Complete Linkage (En Uzak Komşu Yöntemi), iki küme arasındaki mesafeyi, bu kümelerdeki en uzak noktalar arasındaki mesafeyi ölçerek belirler.

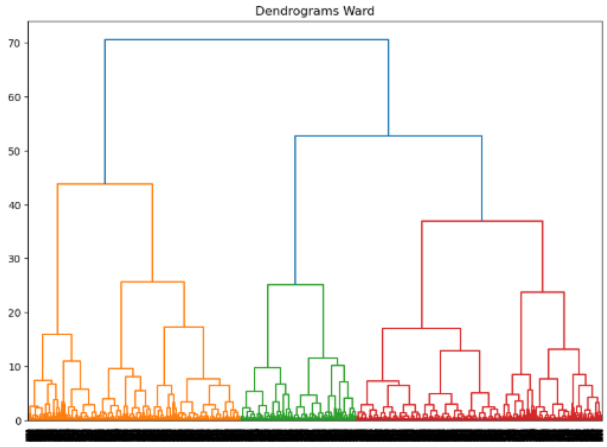
Bu yöntem, kümeler arasında daha sıkı ve kompakt gruplamalar oluşturma eğilimindedir. Çünkü birleşmeyi en uzak mesafeye göre belirler, dolayısıyla uç noktaların kümeleri birleştirme üzerindeki etkisi azalır.

Complete Linkage dendrogramu, Single Linkage'e göre daha düzenli ve net ayrılmış bir yapı sunar. Kümeleme işlemi sırasında veri noktalarının gruplaşması daha tutarlı gözükmektedir.

Genel Yorum:

Acil servise yönelik bu problemde, Complete Linkage yöntemi daha uygun gibi görünüyor. Çünkü sıkı ve daha ayrıştırıcı kümeler, farklı hasta gruplarına yönelik daha kesin ve anlamlı müdahale stratejilerinin belirlenmesine olanak tanır.

Ward yöntemi, kümeler arasındaki toplam kareler toplamını (sum of squares) minimize etmeye çalışır. Bu, her birleşmede kümelerin içsel varyansını en düşük seviyede tutmayı hedefler. Amaç, birbirine benzer veri noktalarını daha sıkı gruplar halinde bir araya getirmek ve iyi ayrılmış kümeler oluşturmaktır.



Dikey eksende büyük boşluklar görülüyor. Bu, veri kümesinin belirgin küme yapısına sahip olduğunu gösterir. Grafiğin en üstündeki büyük birleşimler (örneğin, 60-70 birimlik mesafelerde) kümeleme sonucunda birkaç büyük grubun oluştuğunu işaret eder.

Verilerden daha net ayrışmış hasta grupları elde etmek için Ward yöntemi tercih edilmelidir. Ward yöntemi, farklı hasta profillerinin daha doğru gruplandırılmasına olanak tanır ve acil müdahale gereksinimlerine göre önceliklendirme yapılabilir.

Ward Yöntemi Avantajlı:

- Ward yöntemi, hasta gruplarının belirgin ayrışmasını sağlar. Bu, kritik müdahale gerektiren hasta gruplarını daha kolay belirlemek için uygundur.
- Örneğin, Ward dendrogramında net olarak ayrılan büyük kümeler, hasta gruplarının aciliyet düzeylerine göre (yüksek, orta, düşük) sınıflandırılabilir.

Complete Linkage Yöntemi:

- Complete Linkage yöntemi de dengeli gruplar oluşturma potansiyeline sahiptir. Ancak, hasta gruplarının içsel çeşitliliği fazla ise, bu yöntem bazen aşırı sıkı gruplar oluşturabilir.
- Örneğin, farklı klinik özelliklere sahip hastalar aynı grupta yer alabilir.

Single Linkage Yöntemi:

- Single Linkage yöntemi, zincirleme etkisi nedeniyle acil servis verilerinde tutarsız sonuçlar üretebilir.
- Bu yöntemin yalnızca başlangıç analizlerinde veya kümelerin genel bir fikrini edinmek için kullanılması önerilir.

III.SONUÇ

K-Means++ başlangıç merkezlerini daha iyi seçtiği için, sonuçlar K-Means'e göre daha iyi. K-Means++ rastgele başlangıca göre daha iyi performans göstermiştir. Ancak sonuçlar halen düşük kaliteli kümeler oluşturmaktadır.

PCA analizi varyansın büyük bir kısmını açıklayamıyor (kümülatif varyans düşük). Bu nedenle hiyerarşik yaklaşım denemek daha mantıklı.

Single Linkage kümeleri genellikle tek bir büyük küme oluşturduğu için, acil hastaları belirlemekte başarısız olabilir.

Average Linkage biraz daha iyi çalışsa da, sonuçlar yeterince iyi değil.

Silhouette Score'ların genel olarak düşük olması, kümeler arası ayrışmanın yeterli olmadığını gösteriyor.

Davies-Bouldin Score'un bazı durumlarda yüksek olması, kümelerin birbiriyle fazla iç içe geçtiğini ve belirgin bir ayrım olmadığını gösteriyor.

Acil servise yönelik bu problemde, Complete Linkage yöntemi daha uygun gibi görünüyor. Çünkü sıkı ve daha ayrıştırıcı kümeler, farklı hasta gruplarına yönelik daha kesin ve anlamlı müdahale stratejilerinin belirlenmesine olanak tanır.

KAYNAKÇA

https://medium.com/deep-learning-turkiye/k-means-algoritmas%C4%B1-b460620dd02a

https://www.veribilimiokulu.com/kumeleme-notlari-2-k-ortalamalar/

https://erdincuzun.com/makine_ogrenmesi/hiyerarsik-kumeleme-hierarchical-clustering-odev-benzerlikleri-uzerinden-kopya-gruplarini-bulma/

https://www.displayr.com/what-is-hierarchical-clustering/

https://www.datacamp.com/blog/clustering-in-machine-learning-5-essential-clustering-algorithms