İçindekiler

1. Giriş

Günümüzde büyük ve karmaşık veri setlerinin analizi, bilgiye dayalı karar verme süreçlerinde kritik bir rol oynamaktadır ancak yüksek boyutlu veriler, analiz ve yorumlama açısından zorluklara yol açabilmektedir. Bu tür durumlarda, verinin boyutunu indirgemek ve önemli yapısal özelliklerini öne çıkarmak amacıyla Temel Bileşenler Analizi (Principal Component Analysis - PCA) yaygın olarak kullanılmaktadır. PCA, veri setindeki değişkenler arasındaki ilişkileri dikkate alarak orijinal değişkenlerin doğrusal kombinasyonlarından oluşan yeni bileşenler oluşturur. Böylece veri daha az sayıda bileşenle temsil edilse bile açıklanan toplam varyans korunmaya çalışılır.

Bu çalışmada, UCI Machine Learning Repository' den alınan Seeds veri seti üzerinde Temel Bileşenler Analizi (PCA) uygulanmıştır. Veri setinin yapısı incelenmiştir. Seeds veri seti üç farklı buğday türüne ait toplam yedi farklı fiziksel özellik içermektedir. PCA yardımıyla, veri setindeki yüksek boyutlu özellikler, verideki temel varyans bileşenlerine indirgenmiştir. Bu süreç, verinin özündeki anlamlı yapıları ve potansiyel kümelenme ilişkilerini ortaya çıkarmayı amaçlamaktadır.

PCA uygulamasında kullanılan R kodlarının tamamı "Ekler" bölümünde sunulmuştur.

Bu çalışma ile boyut indirgeme işleminin veri analizi ve yorumlama süreçlerine sağladığı faydalar incelenmiş bileşenler üzerinden anlamlı yapılar tespit edilmeye çalışılmıştır.

2. Temel Bileşen Analizi (PCA) Nedir?

Temel Bileşen Analizi (Principal Component Analysis - PCA), büyük veri kümelerindeki boyut sayısını azaltarak, orijinal bilgilerin büyük kısmını koruyan daha basit ve yorumlanabilir bileşenlere dönüştüren güçlü bir istatistiksel tekniktir. Bu yöntem birbirleriyle potansiyel olarak ilişkili olan değişkenleri, temel bileşenler olarak adlandırılan daha az sayıda yeni değişkene dönüştürür. Bu sayede veriler daha küçük bir uzayda temsil edilerek analizlerin daha hızlı ve etkili bir şekilde gerçekleştirilmesi sağlanır. PCA' nın amacı yalnızca boyut indirgemek değil aynı zamanda verinin altında yatan gizli yapıları keşfetmek ve veri gürültüsünden arınmış anlamlı kalıplar ortaya çıkarmaktır.

PCA' nın kökenleri, 1901 yılında Karl Pearson tarafından geliştirilmiş, modern bilgisayar teknolojilerinin ilerlemesiyle birlikte yaygınlaşarak akademik ve endüstriyel alanlarda önemli bir araç haline gelmiştir. Yüksek boyutlu veri kümelerindeki hareketleri tespit etmekte etkili olan PCA aynı zamanda görselleştirme için boyut indirgeme sağlayarak verinin sezgisel keşfine olanak tanımaktadır. Makine öğrenimi ve veri bilimi uygulamalarında özellikle veri ön işleme aşamasında kullanılan PCA, büyük veri setlerinden en bilgilendirici özellikleri çıkararak büyük boyutlu verilerle çalışmanın zorluğundan kurtarır. Özellikle çoklu doğrusallık ve aşırı uyum (overfitting) gibi problemlerin etkisini azaltmaktadır [1].

Temel Bileşen Analizi (PCA), çok yönlü bir araç olarak çeşitli alanlarda yaygın şekilde kullanılmaktadır. Makine öğrenimi, yapay zekâ, gürültü filtreleme, özellik çıkarımı, güvenilirlik analizi gibi alanlarda yaygın kullanım alanı bulmaktadır. Görüntü işleme ve sinyal işlemede veri sıkıştırma ve analiz kalitesini iyileştirirken gen analizinden genetik verilerdeki gizli yapıları keşfetmek için kullanılmaktadır. Finansal analizde risk değerlendirme ve portföy optimizasyonu sağlarken, pazarlama analitiğinde müşteri davranışlarını anlamaya ve stratejik kararlar geliştirmeye yardımcı olmaktadır. PCA' nın sağladığı avantajlar farklı disiplinlerde kullanılmasını sağlamaktadır [2].

PCA uygulama sürecini 8 aşama altında değerlendirebiliriz. Bu aşamaları adım adım inceleyelim [3].

- **Veri Temizleme**

Veri setinde eksik, tutarsız veya hatalı değerler olup olmadığı kontrol edilir.

- **Standardizasyon**

Değişkenlerin farklı ölçeklerde olması PCA sonuçlarını etkileyebilmektedir. Bu nedenle her bir değişken ortalaması 0 ve varyansı 1 olacak şekilde standardize edilmektedir. Burada amaç tüm değişkenlerin aynı öneme sahip olmasını sağlamaktır. Formül olarak;

$$Z = \frac{X - \mu}{\sigma}$$

Kullanılmaktadır. Z standardize edilmiş değer, X orijinal veri, μ ortalama, σ ise standart sapmadır.

- **Kovaryans veya Korelasyon Matrisi Oluşturulması**

PCA uygulamasında verinin ilişkilerini anlamak için kovaryans veya korelasyon matrisi kullanılmaktadır. Ancak PCA işlemi öncesinde veri standardize edildiğinden genellikle analizde korelasyon matrisi tercih edilmektedir. Korelasyon matrisi değişkenlerin doğrusal ilişkileri gösterir ve bu ilişkiler PCA belirlenmesinde rol oynamaktadır [4].

Kovaryans; iki değişkenin birlikte nasıl değiştiğini ölçmektedir. İki değişkenin ortalamalarına göre pozitif veya negatif yönde ne kadar değişim gösterdiklerinin anlaşılmasını sağlamaktadır[5].

- ✓ Pozitif kovaryans; iki değişken aynı yönde hareket etmektedir(ikisi de artar)
- ✓ Negatif kovaryans; iki değişken ters yönde hareket etmektedir (biri artarken diğeri azalır)
- ✓ 0' a yakın kovaryans; Değişkenler arasında belirgin bir ilişki yoktur

İki değişken (X ve Y) arasındaki kovaryans şu şekilde hesaplanmaktadır.

$$Kovaryans\ (X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

- X_i ve Y_i; X ve Y değişkenlerinin i. gözlemidir.
- $\bar{X}$ ve $\bar{Y}$; X ve Y değişkenlerinin ortalamalarıdır.
- n; gözlem sayısıdır

Kovaryansın birimi iki değişkenin birimlerinin çarpımına bağlıdır. Bu yüzden farklı ölçeklerdeki değişkenlerle çalışırken kovaryans yerine korelasyon tercih edilmektedir.

Korelasyon; İki değişken arasındaki doğrusal ilişkinin yönünü ve gücünü göstermektedir. Kovaryansa göre standartlaştırılmış bir ölçüm sağlamaktadır. Korelasyon katsayısı -1 ile +1 arasında değer almaktadır.

- ✓ +1 : Pozitif doğrusal ilişki
- ✓ -1 : Negatif doğrusal ilişki
- ✓ 0 : Hiçbir ilişki yok

Korelasyon katsayısı kovaryansın her iki değişkenin standart sapmalarına bölünmesiyle elde edilmektedir.

$$\tau_{X,Y} = \frac{Kovaryans\ (X,Y)}{\sigma_X\ \sigma_Y}$$

- ➢ $\tau_{X,Y}$: X ve Y arasındaki korelasyon katsayısı
- ➢ $Kovaryans\ (X,Y)$: X ve Y değişkenlerinin kovaryansı
- ➢ $\sigma_X\ ve\ \sigma_Y$: X ve Y' nin standart sapmaları

Elde edilen bu kovaryans ve korelasyon katsayısı değerleri ile aşağıdaki gibi kovaryans ve korelasyon matrisleri oluşturulmaktadır.

Kovaryans Matrisi;

Bir veri setindeki tüm değişkenlerin birbiriyle olan kovaryanslarını gösteren kare matristir. Köşegen elemanları her değişkenin kendi varyansı ile aynıdır. Örnek 3 değişkenli bir veri seti için kovaryans matrisi şu şekilde gösterilebilir.

$$Kovaryans\ Matrisi = \begin{pmatrix} Cov(X_1,X_1) & Cov(X_1,X_2) & Cov(X_1,X_3) \\ Cov(X_2,X_1) & Cov(X_2,X_2) & Cov(X_2,X_3) \\ Cov(X_3,X_1) & Cov(X_3,X_2) & Cov(X_3,X_3) \end{pmatrix}$$

- ➢ $Cov(X_i,X_j)$: $X_i\ ve\ X_j$ değişkenleri arasındaki kovaryans
- ➢ Köşegen elemanlar : Her değişkenin kendi varyansıdır.

Korelasyon Matrisi;

Veri setindeki her değişkenin diğerleriyle olan korelasyonlarını göstermektedir. Kovaryans matrisi gibi kare matristir. Tüm elemanları -1 ile +1 arasında değer alır. Köşegen elemanları her zaman 1'dir çünkü her değişken kendisiyle tam ilişkilidir. Örnek 3 değişkenli bir veri seti için korelasyon matrisi şu şekilde gösterilebilir.

$$Korelasyon\ Matrisi = \begin{pmatrix} 1 & \tau_{X_1,X_2} & \tau_{X_1,X_3} \\ \tau_{X_2,X_1} & 1 & \tau_{X_2,X_3} \\ \tau_{X_3,X_1} & \tau_{X_3,X_2} & 1 \end{pmatrix}$$

Burada, $\tau_{X_i,X_j}, X_i \; ve \; X_j$ değişkenleri arasındaki korelasyon katsayısını ifade etmektedir.

- **Özdeğer ve Özvektörlerin Hesaplanması**

Elde edilen korelasyon matrisinden özdeğerler ve özvektörler hesaplanmaktadır. PCA analizinde özdeğerler ve özvektörler kilit rol oynamaktadır. Böylece boyut indirgeme mümkün kılınmaktadır ve verinin en çok varyansa sahip yönleri belirlenebilmektedir.

Özvektör, bir matrisin (kovaryans ya da korelasyon matrisi gibi) bir dönüşümü altında yönü değişmeyen bir vektördür ancak yalnızca büyüklüğü değişir. PCA' da her bir özvektör bir temel bileşenin yönünü belirtmektedir.

Özdeğer, özvektörün büyüklüğünün ne kadar değiştiğini ifade eden skaler bir değerdir. PCA' de özdeğer, her bir temel bileşenin veri setindeki varyansa ne kadar katkıda bulunduğunu göstermektdir. Daha yüksek özdeğere sahip bileşenler, verideki daha büyük varyansı açıklamaktadır[6].

Özdeğer-Özvektör hesaplaması şu şekilde yapılabilir;

$$A \, . \, v = \lambda \, . \, v$$

- A : Kovaryans veya korelasyon matrisi
- v : Özvektör (temel bileşenin yönü)
- λ : Özdeğer (temel bileşenin varyansa katkısı)

- **Temel Bileşenlerin Seçilmesi**

Elde edilen özdeğerler azalan sıraya göre sıralanmakta ve varyansı en yüksek derece de açıklayan bileşenler seçilmektedir. Böylece boyut indirgeme gerçekleştirilmektedir.

Scree plot grafiği ile kırılma noktaları belirlenerek uygun bileşen sayısına karar verilmektedir.

Elde edilen özvektörler ile PC1-PC2 gibi temel bileşenler belirlenmektedir.

- **Verinin Yeni Bileşen Uzayına Dönüştürülmesi**

Orijinal veri seti, seçilen temel bileşenler kullanılarak yeni bir bileşen uzayına dönüştürülmektedir. Bu adımda her bir gözlem için yeni bileşen (PC1 – PC2) değerleri hesaplanmaktadır. Standardize edilmiş verinin özvektörlerden oluşan bileşen matrisi ile çarpılması ile bulunmaktadır.

- **Sonuçların Görselleştirilmesi ve Yorumu**

Elde edilen bileşenler kullanılarak veri görselleştirilirmektedir. PC1 ve PC2 eksenlerinde çizilen dağılım grafikleri kümeler arasındaki ayrışmayı göstermektedir.

- **Varyans Açıklama Oranlarının Değerlendirilmesi**

Her bir bileşenin veri setindeki toplam varyansa katkısı (özdeğer oranı) değerlendirilirmektedir.. Kısaca ilk birkaç bileşenin toplam varyansın büyük bir kısmını açıklaması başarılı bir boyut indirgeme yapıldığını göstermektedir.

3. Seeds Veri Seti PCA Uygulaması

Seeds veri setine PCA uygulanarak yedi fiziksel özellik arasındaki varyans yapısı incelenmiş ve verinin bileşen uzayında görselleştirilmesi sağlanmıştır. PCA sonucunda veri setindeki varyansın büyük bir kısmını açıklayan bileşenler elde edilmiş ve farklı buğday türleri arasındaki ayrışma net bir şekilde gözlemlenmiştir. Aşağıda analiz sürecinin aşamaları ve elde edilen bulgular özetlenmiş veri setinde boyut indirgeme sağlanmıştır. Analizin tüm detayları ise "Ekler" bölümünde sunulmuştur.

3.1. Seeds Veri Seti ve Hakkında Bilgi

PCA uygulamasında kullanılan Seeds veri seti, üç farklı buğday türünün fiziksel özelliklerini içieren bir veri setidir. Veri seti, farklı türlerin ayırt edici özelliklerini analiz etmek ve sınıflandırma gibi makine öğrenmesi tekniklerine temel sağlamak amacıyla kullanılmaktadır. UCI Machine Learning Repository' de yayımlanan bu veri seti, çeşitli bilimsel çalışmalarda boyut indirgeme, kümeleme ve sınıflandırma gibi analizlerin uygulanması için sıklıkla tercih edilmektedir.

Seeds veri setinde, üç farklı buğday türü olan Kama, Rosa ve Canadian türlerine ait toplan 210 gözlem bulunmaktadır. Her bir gözlem, buğday tohumlarının fiziksel özelliklerini temsil eden 7 sürekli değişken içermektedir. Bunlar;

- **Area (Alan) :** Tohumun yüzey alanı
- **Perimeter (Çevre) :** Tohumun çevre uzunluğu
- **Compactness (Kompaktlık) :** $\frac{4\pi \times Alan}{Çevre^2}$ formülüyle hesaplanan şekil yoğunluğu
- **Length of Kernel (Çekirdek Uzunluğu) :** Tohum çekirdeğinin uzunluğu
- **Width of Kernel (Çekirdek Genişliği)** : Tohum çekirdeğinin genişliği
- **Asymmetry Coefficient (Asimetri Katsayısı)** : Tohumun asimetriklik derecesi
- **Length of Kernel Groove (Çekirdek Oluk Uzunluğu)** : Çekirdek üzerindeki olukların uzunluğu

Her gözlem için bu özelliklerin yanı sıra buğdayın türünü belirten sınıf etiketi de yer almaktadır. Sınıf etiketleri, veri setinde kategorik olarak 1, 2, ve 3 şeklinde kodlanmıştır;

- **1 :** Kama Buğdayı
- **2 :** Rosa Buğdayı
- **3 :** Canadian Buğdayı

Seeds veri seti, farklı buğday türleri arasındaki yapısal farkları incelemek ve türleri sınıflandırmak amacıyla kullanılmaktadır. Verinin boyut indirgeme yöntemleriyle analizi, bu türlerin birbirinden ne kadar farklı olduğunu anlamaya ve kümeleme yapılarını görselleştirmeye yardımcı olur. Özellikle PCA gibi tekniklerle bu özelliklerin varyans yapısı ortaya çıkarılarak tohum türlerinin ayrışma dereceleri incelenebilmektedir.

Seeds veri seti, sınıflandırma algoritmalarının performansını değerlendirmek için de kullanılabilmektedir. Aynı zamanda boyut indirgeme ve kümeleme gibi denetimsiz makine öğrenme yöntemlerine uygun bir veri setidir[7].

3.2. Seeds Veri Seti Üzerinden PCA Uygulaması (Analiz Süreci)

R programlama dili kullanılarak RStudio ortamında Seeds veri seti üzerinde Temel Bileşenler Analizi (PCA) gerçekleştirilmiştir. Analiz süreci, her bir aşaması detaylı bir şekilde ele alınarak sunulmuştur. İlgili analiz sürecinde kullanılan R kodlarının tamamı Ek-1' de verilmiştir.

R dilinde analizlerin yapılabilmesi için gerekli paketlerin ve kütüphanelerin yüklü olması gerekmektedir. Bu nedenle yapılacak analizlerde kullanılacak paketler ve kütüphaneler analize başlanmadan yüklenmelidir. Yüklenen her bir paketin görevi vardır. Aşağıdaki blokta paketlerin nasıl yükleneceği ve görevleri birlikte verilmiştir.

```
# Paketlerin toplu
yüklenmesi

>
install.packages(c("cluster
",     # Kümeleme
algoritmaları ve analizleri
için

+
"devtools",    # R'de
geliştirme araçları için

+
"pastecs",     #
İstatistiksel özetler ve
tanımlayıcı analizler için

+
"corrplot",    # Korelasyon
matrislerinin
görselleştirilmesi için

+
"factoextra"   # PCA ve
diğer çok değişkenli
analizlerin
görselleştirilmesi için

+ ))
```

Paketlerin Toplu Yüklenmesi

Paketlerin yüklenmesinin ardından bu paketlerin R kodlarında kullanılabilmesi için “library()” komutu ile kütüphanelerinin yüklenmesi gerekmektedir. Aşağıda bu işlem için gerekli blok verilmiştir. Blokta kütüphanelerin görevleri ile birlikte verilmiştir.

Yüklü Paketlerin Kütüphanelerinin Yüklenmesi

Paketlerin ve kütüphanelerin yüklenmesinin ardından analiz sürecine geçmek mümkün hale gelmektedir. RStudio ortamında analizin verimli bir şekilde gerçekleştirilebilmesi için UCI Machine Learning Repository' den indirilen Seeds veri seti ".txt" formatında bilgisayara kaydedilmiştir. Bu veri setine düzenli bir şekilde erişilebilmesi amacıyla " :D " diskinde "seeds_pca" adında özel bir dosya konumu oluşturulmuştur. Bu dosya konumu ayrıca yapılacak analizler için de dosya konumu olacaktır.

" :D" diskinde oluşturduğumuz dosya konumunda bulunan veriyi analize başlayabilmemiz için R ortamına aktarmamız gerekmektedir. Veriyi R ortamına aktarma işlemi aşağıdaki şekilde yapılmıştır. Aşağıda verilen bu blok hakkında açıklama yapmak gerekirse;

read.table() fonksiyonu kullanılarak, belirtilen dosya konumundaki **"seeds_dataset.txt"** dosyası **"data"** adlı bir nesne olarak tanımlanmıştır. Burada **header=FALSE** parametresi, veri setinin ilk satırında değişken isimlerinin yer almadığını belirtmektedir. **sep=""** ifadesi ise veri setindeki sütunların ayrılmasında kullanılan sınırlayıcıyı tanımlamaktadır. Ayrıca, **stringsAsFactors=FALSE** parametresi ile metin verilerinin **faktör** yerine **karakter** tipinde yüklenmesi sağlanmıştır. Bu ayarlar, veri setinin doğru ve tutarlı bir şekilde R ortamına aktarılması için kritik öneme sahiptir. Ayrıca colnames(data) fonksiyonu kullanılarak data nesnesi veri çerçevesindeki sütunlara anlamlı isimler atanmıştır. Anlamlı isimler verilmesi sayesinde veri seti

daha anlaşılır hale getirilmiştir. Anlamlı isimler rastgele verilmemiştir UCI Machine Learning Repository' de bahsedilen parametrelere göre verilmiştir.

```
# TXT dosyasını yükleme
(herhangi bir
sınırlayıcıyla)

data <-
read.table("D:/seeds_pca/se
eds/seeds_dataset.txt",

                    header =
FALSE, sep = "",
stringsAsFactors = FALSE)


# Kolon isimlerini
belirleme (sınıf sütununu
çıkarmak için)

colnames(data) <- c("Area",
"Perimeter", "Compactness",

"Length_of_kernel",
"Width_of_kernel",

"Asymmetry_coef",
"Length_of_kernel_groove",
"Class")
```

.TXT Dosyasını R Ortamına Aktarma

R ortamına aktarılan Seeds veri setinin doğru şekilde aktarılıp aktarılamadığını doğrulamak amacıyla çeşitli kontroller gerçekleştirilmiştir. Yapılan analizlerin sonuçları, veri setine dair bilinen değerler/bilgiler ile karşılaştırılmış ve doğru olduğu tespit edilmiştir. "dim(data)" komutu ile veri setinin boyutu gözlemlenmiştir. "dim(data)" komutu ile veri setinin satır ve sütun sayısı olarak karışlığı elde edilmiştir. "head(data)" komutuyla da verinin ilk 6 satırı gözlenmiştir. Böylece aktarılan veri hakkında bir ön izleme sağlanmıştır.

```
#boyut ve ilk 6 satır

dim(data)

head(data)
```

Aktarılan verinin boyutu ve ilk 6 satırı

R ortamına da verimizi düzgün bir şekilde aktardığımıza göre artık analizlerimize başlayabiliriz. PCA sürecine başlamadan veri setimizin genel yapısı hakkında bilgi sahibi olmak faydalı olacaktır. summary(data) komutu ile veri setinin temel özet istatistiklerini elde edebilmekteyiz. Bu komut bize;

- ✓ **Minimum Değer (Min):** Sütundaki en küçük değer.
- ✓ **Birinci Çeyreklik (1st Qu.):** Gözlemlerin %25'i bu değerin altında yer alır.
- ✓ **Medyan (Median):** Ortanca değer, gözlemlerin %50'si bu değerin altında kalır.
- ✓ **Ortalama (Mean):** Tüm gözlemler toplandıktan sonra gözlem sayısına bölünerek elde edilen aritmetik ortalama.
- ✓ **Üçüncü Çeyreklik (3rd Qu.):** Gözlemlerin %75'i bu değerin altında yer alır.
- ✓ **Maksimum Değer (Max):** Sütundaki en büyük değer.

Bilgilerini vermektedir. Elde ettiğimiz bilgiler sayesinde verinin dağılımı ve merkezi eğilim ölçüleri hızlıca anlaşılabilmektedir. Ayrıca bu bilgiler ışığında çarpıklık ve aykırı değerler tespit edilebilir. Seeds veri setine summary(data) komutu uygulandığında aşağıdaki sonuçlar elde edilmiştir.

```
> summary(data)
      Area         Perimeter      Compactness     Length_of_kernel Width_of_kernel Asymmetry_coef   Length_of_kernel_groove     Class
 Min.   :10.59   Min.   :12.41   Min.   :0.8081   Min.   :4.899    Min.   :2.630   Min.   :0.7651   Min.   :4.519           Min.   :1
 1st Qu.:12.27   1st Qu.:13.45   1st Qu.:0.8569   1st Qu.:5.262    1st Qu.:2.944   1st Qu.:2.5615   1st Qu.:5.045           1st Qu.:1
 Median :14.36   Median :14.32   Median :0.8734   Median :5.524    Median :3.237   Median :3.5990   Median :5.223           Median :2
 Mean   :14.85   Mean   :14.56   Mean   :0.8710   Mean   :5.629    Mean   :3.259   Mean   :3.7002   Mean   :5.408           Mean   :2
 3rd Qu.:17.30   3rd Qu.:15.71   3rd Qu.:0.8878   3rd Qu.:5.980    3rd Qu.:3.562   3rd Qu.:4.7687   3rd Qu.:5.877           3rd Qu.:3
 Max.   :21.18   Max.   :17.25   Max.   :0.9183   Max.   :6.675    Max.   :4.033   Max.   :8.4560   Max.   :6.550           Max.   :3
```

summary(data) komutuyla elde edilen bilgiler

Analizden elde edilen sonuçlara göre verideki her değişkeni şu şekilde yorumlayabilir ve verinin genel yapısını anlayabiliriz. ***Area*** değişkeni geniş bir aralığa yayılmıştır. Ortalama ve medyan birbirine yakın olduğu söylenebilir bu nedenle dağılımın simetrik olduğu kabul edilebilir. Bu durum ölçümlerin dengeli dağıldığını bize göstermektedir. Ortalamadan çok uzak maksimum değer aykırı gözlemlerin bulunabileceğini göstermektedir. ***Perimeter*** değişkeni de simetrik bir dağılıma sahiptir çarpıklık göstermez. Ortalamadan çok uzak maksimum değer aykırı gözlemlerin bulunabileceğini göstermektedir. ***Compactness*** değişkeni çok dar bir aralıkta değişmektedir. Mod ve medyan değerleri birbirine çok yakındır bu nedenle çarpıklık ya da çok uç değerler beklenmemelidir. ***Length_of_kernel*** değişkeninin çok fazla bir değişkenlik gösterdiği söylenemez. Bu nedenle ölçümlerin dengeli olduğu söylenebilir. ***Width_of_kernel*** değişkeninde mod ile medyan birbirine yakındır bu nedenle ölçümün simetrik olduğu söylenebilir. Uç değerlerin sınırlı olduğu söylenebilir. ***Asymetry_coef*** değişkeni geniş bir aralığa yayılmıştır. Buradan aykırı değerler olabileceği anlaşılmaktadır. ***Length_of_kernel_groove*** değişkeninde mod ile medyan birbirine yakındır ve simetrik dağıldığı söylenebilmektedir. Önemli bir aykırı değer varlığı bulunmamaktadır. ***Class*** değişkeni tohumları üç farklı sınıfa ayırır bu nedenle her sınıfın veri setinde dengeli bir şekilde ayrıldığı görülmektedir.

Buradan elde edilen sonuçlarda genel olarak veri setinin dengeli ve simetrik bir dağılıma sahip olduğu görülmektedir. Bazı değişkenlerde aykırı değer beklenmektedir.

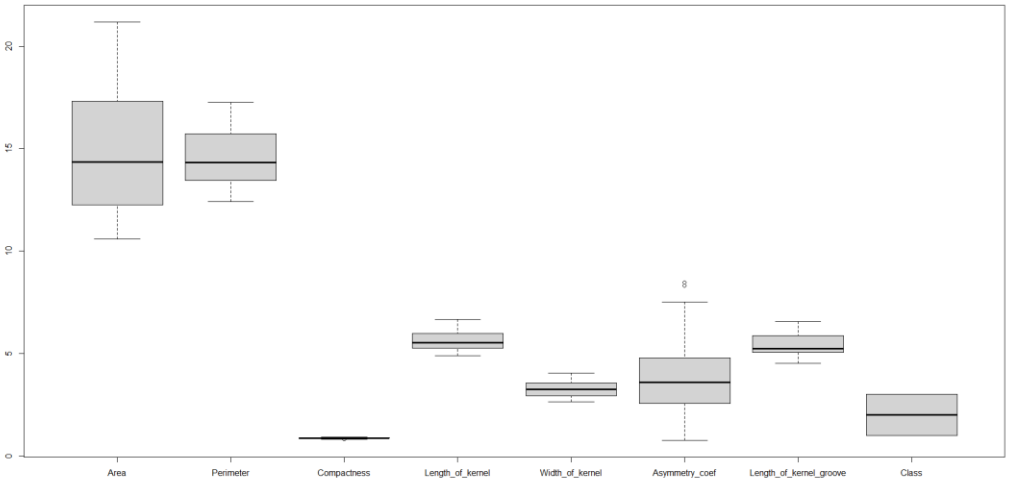
Verinin temel istatistiksel bilgilerinin elde edilmesinin ardından PCA analizine geçilmesi için değişkenlerin standart sapmalarının da elde edilmesi önem göstermektedir. Standart sapma verinin yayılımının anlaşılması, veri standardizasyonu ve PCA gibi yöntemlerde kullanılmaktadır. Bu nedenle verinin standart sapmasının "apply(data,2,sd)" komutuyla elde edilmesi gerekmektedir. Değişkenlerin standart sapmaları aşağıdaki gibi verilmiş ve yorumlanmıştır.

```
apply(data,2,sd)
              Area          Perimeter        Compactness   Length_of_kernel    Width_of_kernel  Asymmetry_coef Length_of_kernel_groove
       2.90969943         1.30595873         0.02362942         0.44306348         0.37771444         1.50355713         0.49148050
             Class
       0.81844759
```

apply(data,2,sd) komutuyla elde edilen standart sapma değerleri

Elde edilen standart sapma değerleri, seeds veri setindeki her bir değişkenin yayılımını-ortalamadan ne kadar uzaklaştığını vermektedir. Bu standart sapma değerleri Area ve Asymmetry_coef değişkenlerinin diğerlerine göre daha yüksek varyansa sahip olduğunu yani bu özelliklerde daha fazla değişiklik olduğunu göstermektedir. Compactness ve Width_of_kernel ise düşük standart sapmaya sahip olup daha homojen bir dağılım görülmektedir.

boxplot(data, horizontal = FALSE) komutu, R' da dikey kutu grafikleri (boxplot) oluşturarak veri setindeki değişkenlerin dağılımını görselleştirmektedir. Bu grafik her bir değişken için merkezi eğilimleri, çeyrekleri ve uç değerleri görsel bir biçimde gözlememizi sağlamaktadır. Seeds veri setindeki tüm verilerin dağılımını, merkezi eğilimlerini ve aykırı değerlerini bizlere veren boxplot aşağıda verilmiştir.



Verisetine ait boxplot grafiği

Boxplot grafiği üzerinden yorum yapmak istersek;

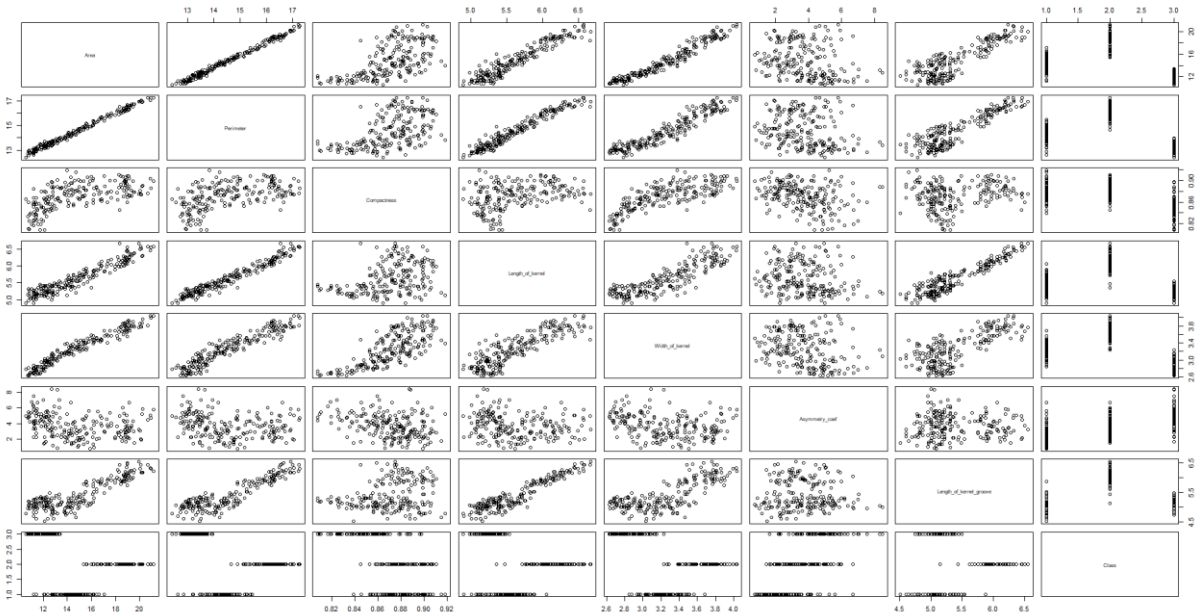
Area değişkeni geniş bir dağılıma sahiptir ve medyan değer kutunun ortasında bulunmaktadır bu durum da simetrik bir dağılım olduğunu göstermektedir. Aykırı değişken yoktur. ***Perimeter*** değişkeni de simetrik bir dağılım göstermekte ancak daha dar bir dağılımı bulunmaktadır. Aykırı değişken yoktur. ***Compactness*** değişkeni oldukça düşük bir yayılıma sahiptir. Bu durumda örneklerin benzer değerlere sahip olduğu söylenebilmektedir. ***Length_of_kernel*** ve ***Width_of_kernel*** değişkenleri orta derecede bir yayılım göstermekte ve aykırı değişken bulundurmamaktadır. Dağılım dengeli ve ortalama etrafında yoğunlaşmıştır. ***Asymmetry_coef*** değişkeni aykırı değerlere sahiptir. Orta derecede yayılım gösterirken üst uçta aykırı değer bulunmaktadır. Asimetri katsayısındaki bu farklılıklar bazı tohumların asimetrik olduğunu göstermektedir. ***Length_of_kernel_groove*** orta derecede yayılım göstermekte ve değerler genel olarak dengeli dağılım sergilemektedir. Aykırı değer yoktur. ***Class*** kategorik bir değişken olduğundan daha az yayılım göstermektedir. Sınıflar arası düzenli bir dağılım vardır.

Yukarıda verilen summary(data) sonuçları verinin temel istatistiklerini sağlar ve veri hakkında genel bir fikir vermektedir. Daha kapsamlı bir istatistiksel özel isteniyorsa stat.desc(data) komutuyla aşağıdaki gibi sonuçlar elde edilebilir.

```
> stat.desc(data)
                      Area    Perimeter  Compactness Length_of_kernel Width_of_kernel Asymmetry_coef Length_of_kernel_groove        Class
nbr.val       210.0000000 2.100000e+02 2.100000e+02     2.100000e+02    210.00000000    210.0000000           2.100000e+02 210.00000000
nbr.null        0.0000000 0.000000e+00 0.000000e+00     0.000000e+00      0.00000000      0.0000000           0.000000e+00   0.00000000
nbr.na          0.0000000 0.000000e+00 0.000000e+00     0.000000e+00      0.00000000      0.0000000           0.000000e+00   0.00000000
min            10.5900000 1.241000e+01 8.081000e-01     4.899000e+00      2.63000000      0.7651000           4.519000e+00   1.00000000
max            21.1800000 1.725000e+01 9.183000e-01     6.675000e+00      4.03300000      8.4560000           6.550000e+00   3.00000000
range          10.5900000 4.840000e+00 1.102000e-01     1.776000e+00      1.40300000      7.6909000           2.031000e+00   2.00000000
sum          3117.9800000 3.057450e+03 1.829097e+02     1.181992e+03    684.30700000    777.0422000           1.135695e+03 420.00000000
median         14.3550000 1.432000e+01 8.734500e-01     5.523500e+00      3.23700000      3.5990000           5.223000e+00   2.00000000
mean           14.8475238 1.455929e+01 8.709986e-01     5.628533e+00      3.25860476      3.7002010           5.408071e+00   2.00000000
SE.mean         0.2007883 9.011971e-02 1.630585e-03     3.057428e-02      0.02606477      0.1037553           3.391538e-02   0.05647825
CI.mean.0.95    0.3958300 1.776602e-01 3.214501e-03     6.027352e-02      0.05138356      0.2045411           6.686008e-02   0.11134006
var             8.4663508 1.705528e+00 5.583493e-04     1.963052e-01      0.14266820      2.2606840           2.415531e-01   0.66985646
std.dev         2.9096994 1.305959e+00 2.362942e-02     4.430635e-01      0.37771444      1.5035571           4.914805e-01   0.81844759
coef.var        0.1959720 8.969937e-02 2.712911e-02     7.871739e-02      0.11591294      0.4063447           9.087907e-02   0.40922380
```

Stat.desc(data) sonuçları

Analize geçmeden verimiz hakkında bilgilerin büyük birçoğunu edindik. Analize başlamadan önceki son aşamada veri setindeki değişkenlerin ilişkisini görmek amacıyla pairs(data) komutu kullanılacaktır. Böylece veri setindeki değişkenler arasındaki ilişkiler scatter plot (dağılım grafiği) oluşturularak görselleştirilecektir. Buradan veri setimiz hakkında yorum yapabiliriz.



Veri Seti Scatter Plot

Yukarıda verilen grafiği yorumlamak gerekirse;

Scatter plot üzerinden güçlü doğrusal ilişki, orta düzey ilişki ve zayıf-belirsiz ilişki bulunan durumlar yorumlanabilmektedir. ***Area*** ve ***Perimeter***, ***Length_of_kernel*** ve ***Width_of_kernel***, ***Length_of_kernel*** ve ***Length_of_kernel_groove*** arasında güçlü pozitif doğrusal ilişki bulunmaktadır. Yani alan arttıkça çevre artar, uzun tohumlar daha geniştir ve tohum uzunluğu arttıkça oluk uzunluğu da artmaktadır. ***Compactness*** değişkeni diğer değişkenlerle belirgin bir doğrusal ilişki göstermezken, ***area*** ve ***perimeter*** ile orta derecede ilişki göstermektedir. ***Class*** değişkeni kategorik durumundan dolayı belirsiz bir ilişki göstermektedir.

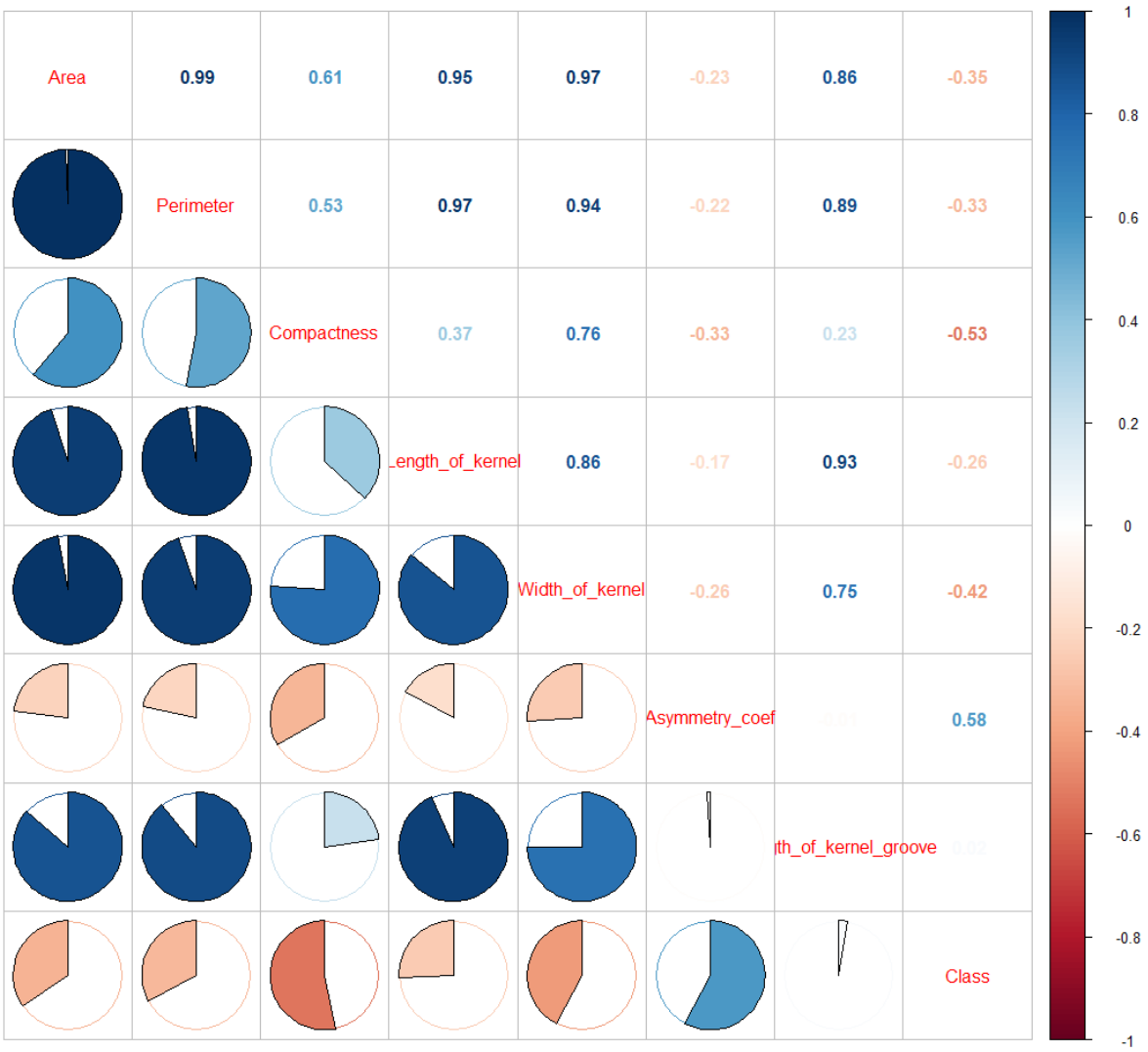
Veri setimizin yapısı hakkında genel bilgilere sahip olduk. Buradan sonra artık PCA aşamalarına başlayabiliriz ve farklı yöntemlerle yapılışını inceleyebiliriz. Bu aşamadan farklı yöntemler ve PCA sonuçlarının bulunuşuyla birlikte analizi verilmiştir.

İlk olarak ***korelasyon matrisi ve özdeğer/özvektörlerin hesaplanması*** ile temel bileşenlerin manual olarak oluşturulması ele alınmıştır. Bu işlem için gerekli korelasyon matrisi aşağıdaki gibi verilmiştir.

```
> corr
                              Area  Perimeter Compactness Length_of_kernel Width_of_kernel Asymmetry_coef Length_of_kernel_groove       Class
Area                     1.0000000  0.9943409   0.6082884        0.9499854       0.9707706    -0.22957233              0.86369275 -0.34605787
Perimeter                0.9943409  1.0000000   0.5292436        0.9724223       0.9448294    -0.21734037              0.89078390 -0.32789970
Compactness              0.6082884  0.5292436   1.0000000        0.3679151       0.7616345    -0.33147087              0.22682482 -0.53100702
Length_of_kernel         0.9499854  0.9724223   0.3679151        1.0000000       0.8604149    -0.17156243              0.93280609 -0.25726870
Width_of_kernel          0.9707706  0.9448294   0.7616345        0.8604149       1.0000000    -0.25803655              0.74913147 -0.42346287
Asymmetry_coef          -0.2295723 -0.2173404  -0.3314709       -0.1715624      -0.2580365     1.00000000             -0.01107902  0.57727271
Length_of_kernel_groove  0.8636927  0.8907839   0.2268248        0.9328061       0.7491315    -0.01107902              1.00000000  0.02430104
Class                   -0.3460579 -0.3278997  -0.5310070       -0.2572687      -0.4234629     0.57727271              0.02430104  1.00000000
```

Seeds veri seti için korelasyon matrisi

Korelasyon matrisinden değerler -1 ile 1 arasında yer almaktadır. 1 mükemmel pozitif korelasyonu, -1 mükemmel negatif korelasyonu, 0 ise ilişki yok durumunu ifade etmektedir. Korelasyon matrisi bu şekilde yorumlanabilir. Korelasyon matrisi dağılımı corrplot.mixed(corr, lower="pie",upper="number") komutuyla aşağıdaki gibi görselleştirilebilir.



Seeds veri seti için korelasyon matrisinin daire grafiği görselleştirilmesi

Hesaplanan korelasyon matrisi, özdeğer ve özvektörlerin hesaplanması için gereklidir. Korelasyon matrisi üzerinden özdeğer ve özvektörlerin hesaplanması gerçekleştirilmektedir. Özdeğer – özvektörlerin elde edilmesiyle veri setinin bileşenlerinin önem derecesi elde edilir. Bu sayede veri setindeki varyansı en iyi açıklayan bileşenler bulunmaktadır. Böylece bilgi kaybı minimize edilerek boyut azaltılmış olur ve veri seti düşük boyutta temsil edilebilir. R üzerinde seeds veri setinin elde edilen özdeğer ve özvektör değerleri aşağıdaki gibidir;

```
$values
[1] 5.1871360642 1.6978805370 0.6790516245 0.3694409739 0.0451347860 0.0153349541 0.0053308498 0.0006902105

$vectors
           [,1]        [,2]        [,3]        [,4]        [,5]         [,6]        [,7]         [,8]
[1,] -0.4354838  0.08082488  0.02699035  0.01006832  0.26132543 -0.121214994 -0.42830865  0.732501880
[2,] -0.4316457  0.11727215 -0.06274185 -0.06446452  0.30465410 -0.135134440 -0.47459965 -0.674919831
[3,] -0.2869907 -0.35972327  0.65905286  0.37492868 -0.32810705  0.280923693 -0.14252979 -0.080163678
[4,] -0.4110251  0.20138426 -0.22206055 -0.17295887  0.05778331  0.797196385  0.26881355  0.012464589
[5,] -0.4282862 -0.02983049  0.22463489  0.08927485  0.32627958 -0.392042277  0.70440081 -0.030236204
[6,]  0.1396575  0.55560659  0.63983095 -0.51157800 -0.01699613  0.006569901 -0.01918675  0.001110252
[7,] -0.3646543  0.38186129 -0.22413150  0.05457218 -0.76052907 -0.295652684  0.04667083 -0.001730277
[8,]  0.1918971  0.59527434  0.03992503  0.74336835  0.20676085  0.107026124 -0.00214970 -0.020595693
```

Seeds veriseti özdeğer ve özvektör değerleri

Özdeğerlere bakıldığında ilk üç bileşen veri setindeki varyansın büyük bir kısmını (%93) açıklamaktadır. Bu durum boyut indirgeme açısından avantaj sağlamaktadır. İlk iki bileşen veri setindeki varyansın %93' ünü koruduğu için yapılan analizlerde yeterli bilgiyi sağlamaktadır. Bu durum PCA' nın asıl amacı yani boyut indirgemeyi başarıyla yerine getirdiğimizi göstermektedir.

Özvektör verileri, analizde her bir temel bileşenin yönünü ve hangi değişkenlerin bu bileşenlere ne kadar katkı sağladığını göstermektedir. Her sütün bir temel bileşeni (PCx) temsil etmekte ve her bir bileşen veri setindeki varyansı belirlemektedir. PC1-PC2-PC3 bileşenlerinde çeşitli değişkenlerde yüksek değerler görülmekte ve varyansı yüksek düzeyde açıkladığı anlaşılmaktadır.

Genel olarak özdeğer ve özvektörlere bakıldığında ilk üç temel bileşenin (PC1-PC2-PC3) veri setindeki varyansın büyük bir kısmını açıkladığı görülmektedir. Bu üç bileşen varyansı açıklayarak boyut indirgemede önemli rol oynamaktadır.

```
                [,1]      [,2]       [,3]       [,4]        [,5]        [,6]         [,7]         [,8]
eigenvalues 5.187136 1.6978805 0.67905162 0.36944097 0.045134786 0.015334954 0.0053308498 6.902105e-04
prop.var    0.648392 0.2122351 0.08488145 0.04618012 0.005641848 0.001916869 0.0006663562 8.627631e-05
cum.prop.var 0.648392 0.8606271 0.94550853 0.99168865 0.997330498 0.999247367 0.9999137237 1.000000e+00
```

Özdeğer bileşenleri hakkında sayısal veriler

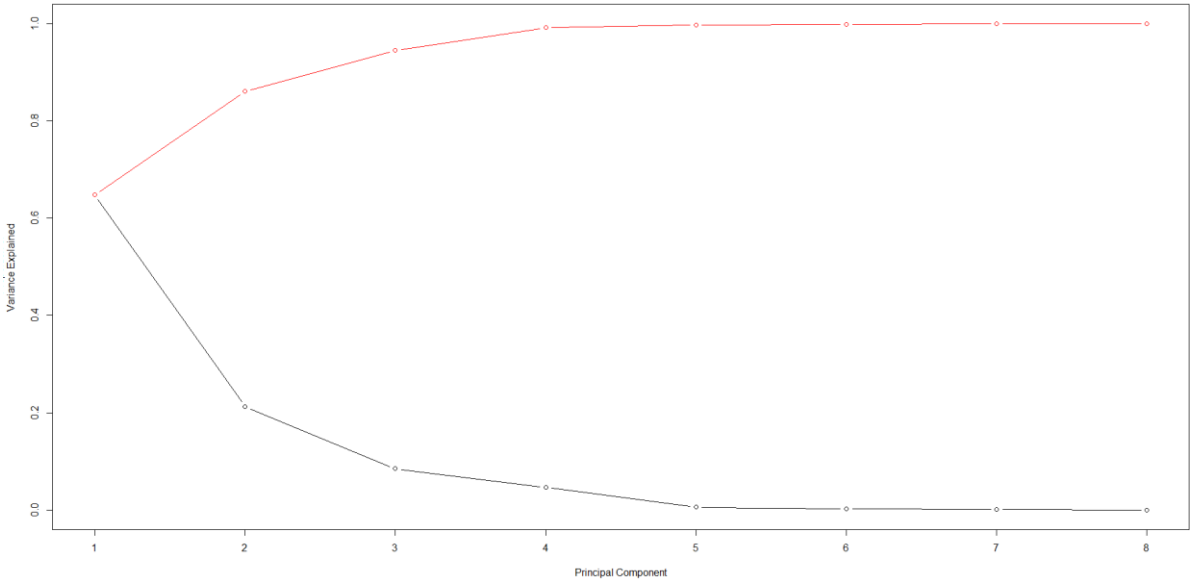
Görüldüğü üzere 8 adet özdeğer temel bileşeni bulunmaktadır. Bu tablodan her bir bileşenin tek başına toplam varyansın ne kadarını açıkladığını gösteren prop.var değeri ve her bir bileşeni ekledikçe toplam varyansın ne kadarının açıklandığını gösteren cum.prop.var değerleri hakkında yorum yapılabilmektedir.

Yüksek prop.var oranına sahip bileşenler, veri setindeki varyansın büyük kısmını açıklamaktadır ve analizde önemli yer tutmaktadır. Boyut indirgeme açısından en varyansı en yüksek açıklayan bileşenlerin seçilmesi gerekmektedir. İlk üç bileşenin %93 civarında açıkladığı görülmektedir. Yüksek değere sahip bileşenlerin alınması bilgi kaybını minimize etmektedir. Geri kalan bileşenler göz ardı edilebilir.

Cum.prop.var oranlarına bakıldığında ilk bileşenin %65 civarında varyansı açıkladığı ve bilginin büyük bir kısmının bu bileşende temsil edildiği görülmektedir. Kümülatif olarak bakıldığında ilk iki bileşen varyansın yaklaşık

%87 sini açıklamaktadır. Bu iki bileşen kümülatif olarak büyük bilgi açıklamaktadır. %90 sınırını aşmak için üçüncü bileşende dahil edilebilir. Geri kalan bileşenler varyansın az bir miktarını açıkladığı için göz ardı edilebilir.

Açıklanan prop.var ve cum.prop.var değerleri tek bir grafik üzerinde gösterilebilmektedir. "***plot(prop.var)***" komutu ile her bir bileşenin açıkladığı varyans oranını gösteren bir grafik çizilmektedir. Bu grafikte X ekseni bileşenleri (PC1, PC2, …) Y ekseni ise her bir bileşenin açıkladığı varyans oranını göstermektedir. "***lines(cum.prop.var, type='b', col='red')***" komutuyla kümülatif varyans oranı kırmızı renkte çizgi ile gösterilmektedir. Bileşenler eklendikçe toplam varyansın ne kadarınnı açıklandığı görsel olarak ifade edilmektedir. Aşağıdaki grafikte değerler görselleştirmiştir.



Varyans – kümülatif varyans , temel bileşen grafiği

Bu grafikte siyah çizgi varyansı, kırmızı çizgi kümülatif varyansı ifade etmektedir. Grafikten varyans açıklama oranının ilk bileşende en yüksek olup diğer bileşenlere doğru düştüğü görülmektedir. Kümülatif varyansı ilk üç bileşende %90 civarına yaklaştığı ve diğer bileşenlerin katkısının minimum olduğu görülmektedir. Bu grafik, boyut indirgeme işlemi için kaç bileşenin yeterli olacağına karar vermeyi kolaylaştırmaktadır.

Korelasyon matrisi üzerinden gerçekleştirilen PCA analiziyle özdeğer ve özvektörlerin hesaplanması sonucunda temel bileşenler elde edilmiştir. Elde edilen her bir temel bileşenin veri setindeki toplam varyanstan ne kadarını açıkladığı ve toplam varyansın kümülatif olarak ne kadarını kapsadığı gösterilmiştir.

Temel bileşenler bu şekilde manual olarak hesaplanabileceği gibi ayrıca ***paket üzerinden*** de kolayca hesaplanabilmektedir.

prcomp fonksiyonu sayesinde PCA otomatik olarak hesaplanmaktadır. Bir önceki aşamada sırayla hesaplanan korelasyon matrisinin bulunması ve ardından özdeğer-özvektör hesaplanması işlemlerini bu fonksiyon otomatik olarak gerçekleştirmektedir. Paket fonksiyon ile elde edilen sonuçlar aşağıda verilmiştir. İlgili R kodları ek-1'de verilmiştir.

```
> data.pca <- prcomp(data, center = TRUE, scale. = TRUE)  ##korelasyon matrisi üzerinden
> summary(data.pca)
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7     PC8
Standard deviation     2.2775 1.3030 0.82405 0.60782 0.21245 0.12383 0.07301 0.02627
Proportion of Variance 0.6484 0.2122 0.08488 0.04618 0.00564 0.00192 0.00067 0.00009
Cumulative Proportion  0.6484 0.8606 0.94551 0.99169 0.99733 0.99925 0.99991 1.00000
```

Korelasyon matrisi üzerinden fonksiyon aracılığıyla temel bileşenlerin bulunması

```
> (data.pca$sdev)^2 ##özdeğerler
[1] 5.1871360642 1.6978805370 0.6790516245 0.3694409739 0.0451347860 0.0153349541 0.0053308498 0.0006902105
```

Seeds Veri setinin özdeğerleri

```
> cor(data)
                              Area  Perimeter Compactness Length_of_kernel Width_of_kernel Asymmetry_coef Length_of_kernel_groove       Class
Area                     1.0000000  0.9943409   0.6082884        0.9499854       0.9707706    -0.22957233              0.86369275 -0.34605787
Perimeter                0.9943409  1.0000000   0.5292436        0.9724223       0.9448294    -0.21734037              0.89078390 -0.32789970
Compactness              0.6082884  0.5292436   1.0000000        0.3679151       0.7616345    -0.33147087              0.22682482 -0.53100702
Length_of_kernel         0.9499854  0.9724223   0.3679151        1.0000000       0.8604149    -0.17156243              0.93280609 -0.25726870
Width_of_kernel          0.9707706  0.9448294   0.7616345        0.8604149       1.0000000    -0.25803655              0.74913147 -0.42346287
Asymmetry_coef          -0.2295723 -0.2173404  -0.3314709       -0.1715624      -0.2580365     1.00000000             -0.01107902  0.57727271
Length_of_kernel_groove  0.8636927  0.8907839   0.2268248        0.9328061       0.7491315    -0.01107902              1.00000000  0.02430104
Class                   -0.3460579 -0.3278997  -0.5310070       -0.2572687      -0.4234629     0.57727271              0.02430104  1.00000000
> data.pca$rotation  ##özvektörleri verir.
                                PC1         PC2         PC3         PC4         PC5          PC6         PC7          PC8
Area                    -0.4354838 -0.08082488 -0.02699035  0.01006832 -0.26132543  0.121214994 -0.42830865 -0.732501880
Perimeter               -0.4316457 -0.11727215  0.06274185 -0.06446452 -0.30465410  0.135134440 -0.47459965  0.674919831
Compactness             -0.2869907  0.35972327 -0.65905286  0.37492868  0.32810705 -0.280923693 -0.14252979  0.080163678
Length_of_kernel        -0.4110251 -0.20138426  0.22206055 -0.17295887 -0.05778331 -0.797196385  0.26881355 -0.012464589
Width_of_kernel         -0.4282862  0.02983049 -0.22463489  0.08927485 -0.32627958  0.392042277  0.70440081  0.030236204
Asymmetry_coef           0.1396575 -0.55560659 -0.63983095 -0.51157800  0.01699613 -0.006569901 -0.01918675 -0.001110252
Length_of_kernel_groove -0.3646543 -0.38186129  0.22413150  0.05457218  0.76052907  0.295652684  0.04667083  0.001730277
Class                    0.1918971 -0.59527434 -0.03992503  0.74336835 -0.20676085 -0.107026124 -0.00214970  0.020595693
> data.pca$center  ##değişkenlerin ortalamaları
                    Area                Perimeter              Compactness         Length_of_kernel          Width_of_kernel       Asymmetry_coef Length_of_kernel_groove
              14.8475238               14.5592857                0.8709986                5.6285333                3.2586048                3.7002010                5.4080714
                   Class
               2.0000000
> data.pca$scale   ##değişkenlerin standart sapmaları
                    Area                Perimeter              Compactness         Length_of_kernel          Width_of_kernel       Asymmetry_coef Length_of_kernel_groove
              2.90969943               1.30595873                0.02362942               0.44306348               0.37771444               1.50355713               0.49148050
                   Class
              0.81844759
```

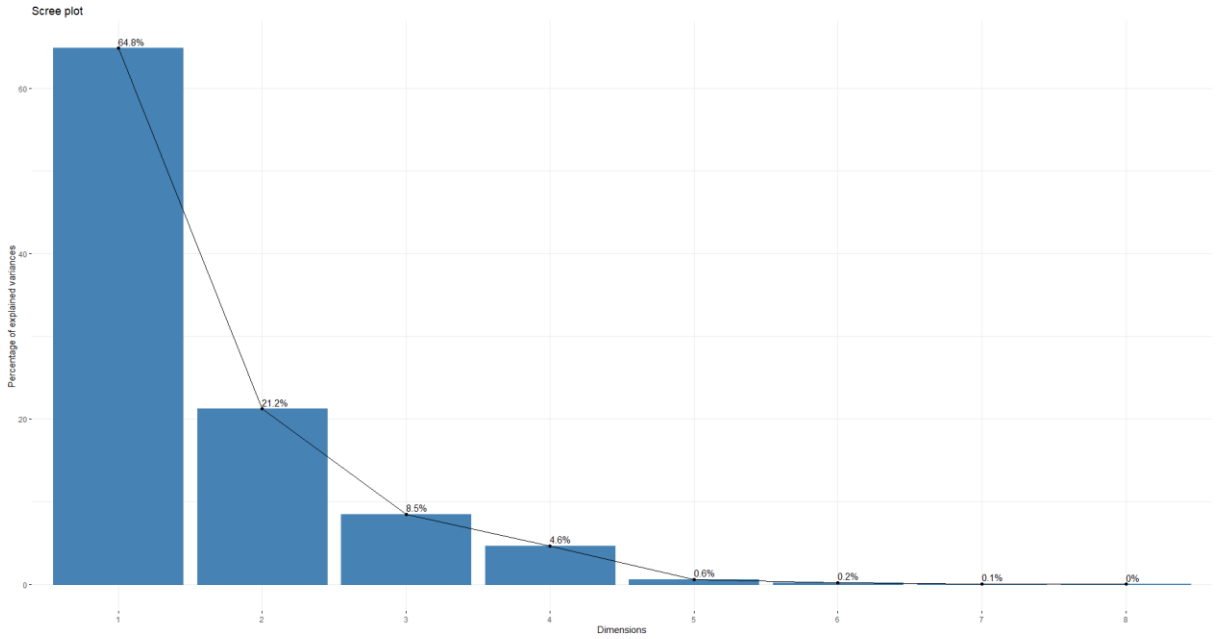
Paket Üzerinden Korelasyon Matrisi, özvektörler, değişken ortalamaları ve standart sapma elde edilmesi

PCA analizi, hem manuel hesaplamalarla hem de R paketleri kullanılarak gerçekleştirilebilmektedir ve her iki yöntemle de temel bileşenler aynı sonuçlarla elde edilebilmektedir. Manuel yöntem, sürecin her adımını ayrıntılı olarak görmemizi sağlarken, paket kullanımı ise daha hızlı ve pratik bir çözüm sunmaktadır. Her iki yöntemin sonuçlarının örtüşmesi, analizimizin doğruluğunu ve tutarlılığını pekiştirmektedir.

3.3. Analizin Görselleştirilmesi

"factoextra" paketi ile PCA gibi çok değişkenli analizlerin sonuçlarını daha anlaşılır ve etkili bir şekilde görselleştirmek için kullanılan bir R paketidir. Bu paket, analiz sonuçlarını grafiklerle ifade ederek bileşenlerin açıklama gücünü net bir biçimde sunmaktadır ve veri içindeki ilişkilerin yorumlanmasını sağlamaktadır. Bu nedenle veri setinin daha kolay yorumlanabilmesi için analiz sonuçları "factoextra" paketi ile görselleştirilmiştir.

İlk olarak "***fviz_eig(data.pca, addlabels = TRUE)***" komutuyla PCA analizinde elde edilen bileşenlerin açıkladığı varyans oranı görselleştirilmiştir. Bu fonksiyon PCA' nın açıklanan varyans oranlarını sütun grafiğinde yüzde değer olarak göstermektedir. Bu grafik "Scree Plot" olarak adlandırılmaktadır. Grafikte ilk bileşenler yüksek varyans, diğer bileşenler düşük varyansa sahiptir ve buradan kaç bileşenin boyut indirgemede yeterli olacağı anlaşılabilmektedir.



Scree Plot

PCA sonuçlarının görselleştirilmesi amacıyla ***"res.var$cos2"*** komutu kullanılarak her bir değişkenin PCA bileşenlerindeki temsil kalitesi hesaplanmıştır. ***"cos2"*** değerleri her değişkenin belirli bir bileşende ne kadar iyi temsil edildiğini ifade etmektedir. Yüksek ***"cos2"*** değerleri değişkenin o bileşende güçlü bir şekilde aldığını göstermektedir. Bu temsil kalitesini daha iyi anlayabilmek amacıyla ***"corrplot(res.var$cos2,is.corr=FALSE)"*** komutu kullanılarak ısı haritası oluşturulmuştur. Bu sayede değişkenlerin bileşenlerdeki etkisi anlaşılabilmektedir.

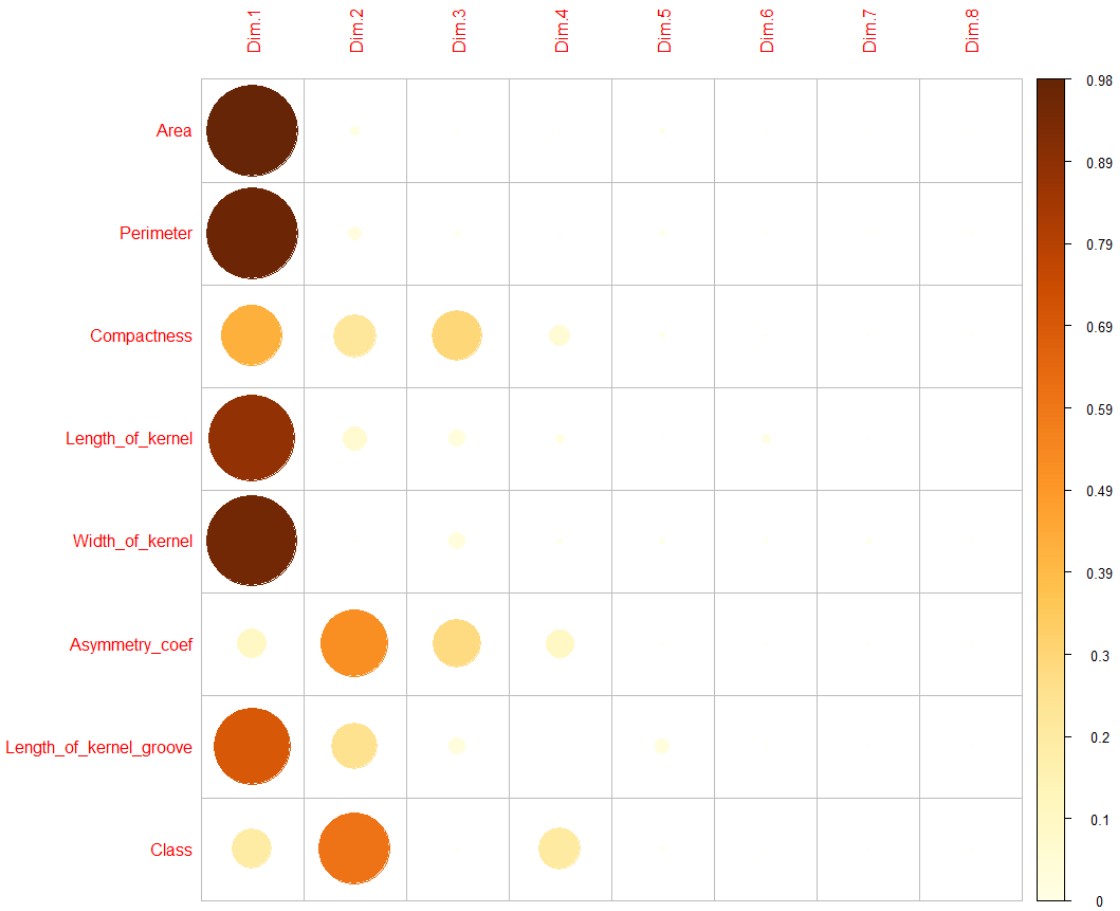
```
> eig.val
       eigenvalue variance.percent cumulative.variance.percent
Dim.1 5.1871360642     64.839200802                    64.83920
Dim.2 1.6978805370     21.223506713                    86.06271
Dim.3 0.6790516245      8.488145306                    94.55085
Dim.4 0.3694409739      4.618012174                    99.16886
Dim.5 0.0451347860      0.564184825                    99.73305
Dim.6 0.0153349541      0.191686926                    99.92474
Dim.7 0.0053308498      0.066635623                    99.99137
Dim.8 0.0006902105      0.008627631                   100.00000
> res.var <- get_pca_var(data.pca)
> res.var$coord          # Coordinates
                             Dim.1       Dim.2       Dim.3       Dim.4       Dim.5         Dim.6         Dim.7         Dim.8
Area                   -0.9918268 -0.10531704 -0.02224129  0.00611969 -0.05551845  0.0150105838 -0.0312719573 -1.924419e-02
Perimeter              -0.9830853 -0.15280884  0.05170216 -0.03918261 -0.06472361  0.0167342899 -0.0346517868  1.773140e-02
Compactness            -0.6536295  0.46872930 -0.54308980  0.22788786  0.06970617 -0.0347880118 -0.0104064803  2.106049e-03
Length_of_kernel       -0.9361214 -0.26240922  0.18298808 -0.10512727 -0.01227603 -0.0987203214  0.0196267946 -3.274679e-04
Width_of_kernel        -0.9754340  0.03886994 -0.18510945  0.05426273 -0.06931793  0.0485483129  0.0514301828  7.943612e-04
Asymmetry_coef          0.3180740 -0.72397064 -0.52725007 -0.31094558  0.00361082 -0.0008135796 -0.0014008756 -2.916837e-05
Length_of_kernel_groove -0.8305106 -0.49757575  0.18469464  0.03316987  0.16157401  0.0366119673  0.0034075619  4.545760e-05
Class                   0.4370512 -0.77565881 -0.03290006  0.45183159 -0.04392624 -0.0132535139 -0.0001569554  5.410871e-04
> res.var$contrib        # Contributions to the PCs
                             Dim.1       Dim.2      Dim.3      Dim.4       Dim.5       Dim.6        Dim.7        Dim.8
Area                    18.964613  0.65326620  0.0728479  0.0101371  6.82909787  1.46930747 1.834483e+01 5.365590e+01
Perimeter               18.631798  1.37527581  0.3936540  0.4155675  9.28141228  1.82613170 2.252448e+01 4.555168e+01
Compactness              8.236367 12.94008310 43.4350676 14.0571515 10.76542330  7.89181213 2.031474e+00 6.426215e-01
Length_of_kernel        16.894165  4.05556195  4.9310887  2.9914771  0.33389108 63.55220761 7.226072e+00 1.553660e-02
Width_of_kernel         18.342907  0.08898579  5.0460833  0.7969998 10.64583671 15.36971468 4.961805e+01 9.142280e-02
Asymmetry_coef           1.950422 30.86986811 40.9383650 26.1712048  0.02888685  0.00431636 3.681313e-02 1.232659e-04
Length_of_kernel_groove 13.297276 14.58180474  5.0234927  0.2978122 57.84044703  8.74105093 2.178166e-01 2.993860e-04
Class                    3.682451 35.43515430  0.1594008 55.2596499  4.27500488  1.14545913 4.621211e-04 4.241826e-02
```

Özdeğerler, konumlar, bileşenleri açıklama değeri

```
> res.var$cos2           # Quality of representation
                             Dim.1       Dim.2        Dim.3        Dim.4        Dim.5        Dim.6        Dim.7        Dim.8
Area                    0.9837203 0.011091680 0.0004946749 3.745061e-05 3.082299e-03 2.253176e-04 9.779353e-04 3.703387e-04
Perimeter               0.9664567 0.023350540 0.0026731137 1.535277e-03 4.189146e-03 2.800365e-04 1.200746e-03 3.144025e-04
Compactness             0.4272316 0.219707152 0.2949465322 5.193288e-02 4.858951e-03 1.210206e-03 1.082948e-04 4.435441e-06
Length_of_kernel        0.8763233 0.068858597 0.0334846381 1.105174e-02 1.507010e-04 9.745702e-03 3.852111e-04 1.072352e-07
Width_of_kernel         0.9514716 0.001510872 0.0342655103 2.944444e-03 4.804976e-03 2.356939e-03 2.645064e-03 6.310098e-07
Asymmetry_coef          0.1011711 0.524133482 0.2779926324 9.668715e-02 1.303802e-05 6.619118e-07 1.962453e-06 8.507939e-10
Length_of_kernel_groove 0.6897478 0.247581625 0.0341121090 1.100240e-03 2.610616e-02 1.340436e-03 1.161148e-05 2.066393e-09
Class                   0.1910137 0.601646588 0.0010824138 2.041518e-01 1.929514e-03 1.756556e-04 2.463498e-08 2.927753e-07
```

Temsil Kalitesi Değerleri

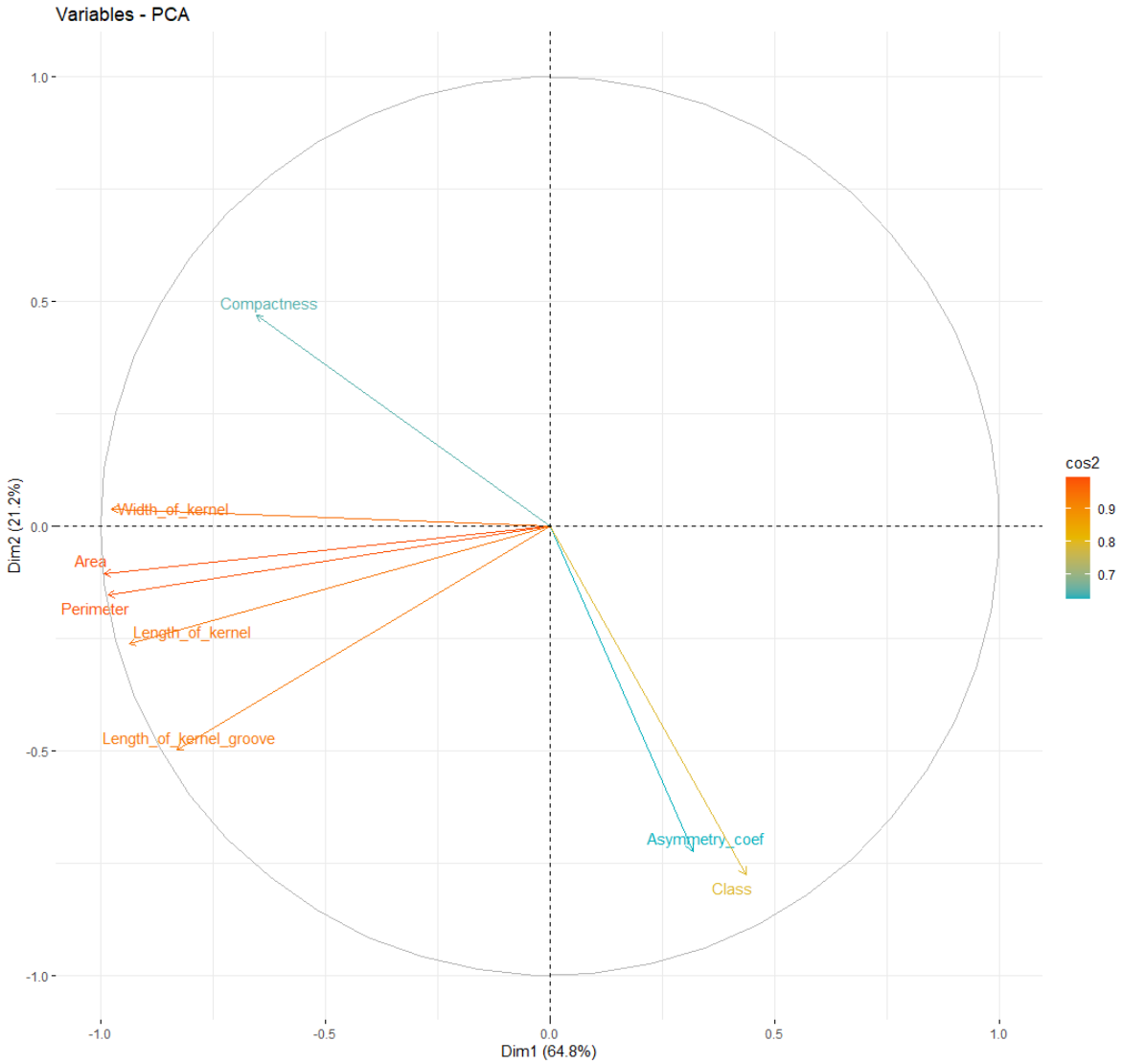
Görülen cos2 değerleri incelendiğinde temel bileşenlerin hangi değişkenler üzerinde daha baskın olduğu, varyansın daha büyük kısmını açıkladığı anlaşılmaktadır. Bu değerler aşağıda heatmap üzerinde görselleştirilmiştir.



Temsil Kalitesi HeatMap

Bu heatmap'ta dairelerin büyüklüğü ve renk yoğunluğu her bileşenin sahip olduğu "cos2" değerinin yansıtmaktadır. Koyu renk ve büyük daireler yüksek "cos2" değerlerini yani güçlü temsili ifade etmektedir. Böylece temel bileşenlerin hangi değişkenler üzerinde güçlü olduğu görsel olarak netleşmektedir.

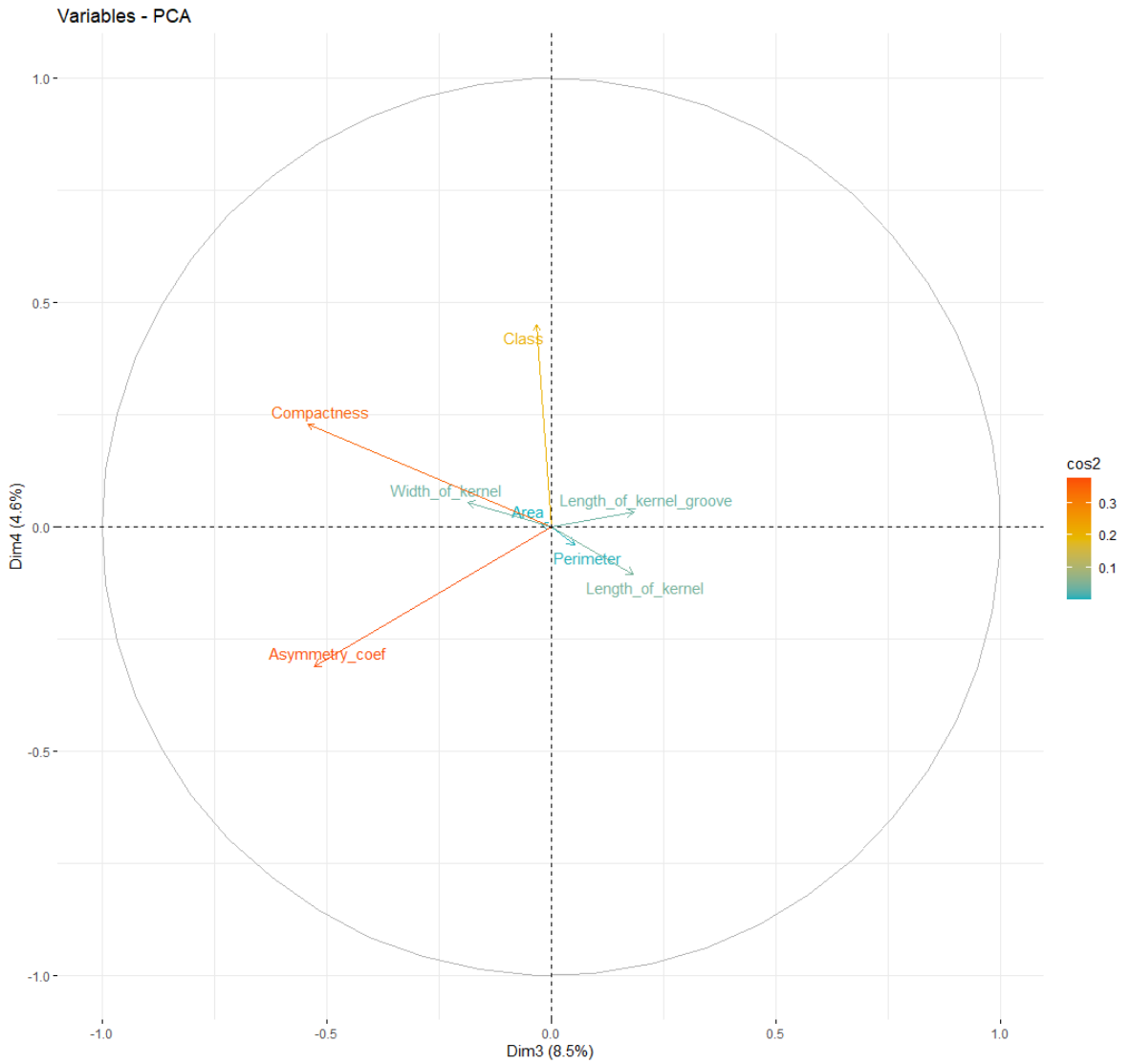
PCA bileşenlerine yapılan değişken katkılarını ve değişkenlerin bileşenler üzerindeki dağılımını görselleştirmek için "fviz_contrib" ve "fviz_pca_var" fonksiyonları kullanılmaktadır. İlk iki temel bileşene en çok katkı sağlayan değişkenler belirlenmiş ve "cos2" değerleriyle temsil kaliteleri renk gradyanı ile gösterilmiştir. Böylece hangi değişkenlerin bileşenlerde baskı olduğu ve bileşenlerde nasıl dağıldığı anlaşılabilmektedir.



İlk iki bileşen üzerinde değişkenler grafiği PC1-PC2

Grafikte, PCA analizinde değişkenlerin ilk iki bileşende nasıl dağıldığı ve bileşenlere katkıları görülmektedir. ***"area"***, ***"perimeter"***, ***"width_of_kernel"*** ve ***"Length_of_kernel"*** gibi değişkenler Dim.1 üzerinde güçlü bir etkiye sahiptir ve birbirleriyle pozitif ilişkidir. ***"Compactness***" ise Dim.2 üzerinde baskın olduğu diğer değişkenlerden farklı bir yöndedir. Renk tonları değişkenlerin bileşenlerdeki temsil kalitesini ***"cos2"*** göstermektedir. Daha koyu tonlar, değişkenin o bileşende daha iyi temsil edildiğini ifade etmektedir.

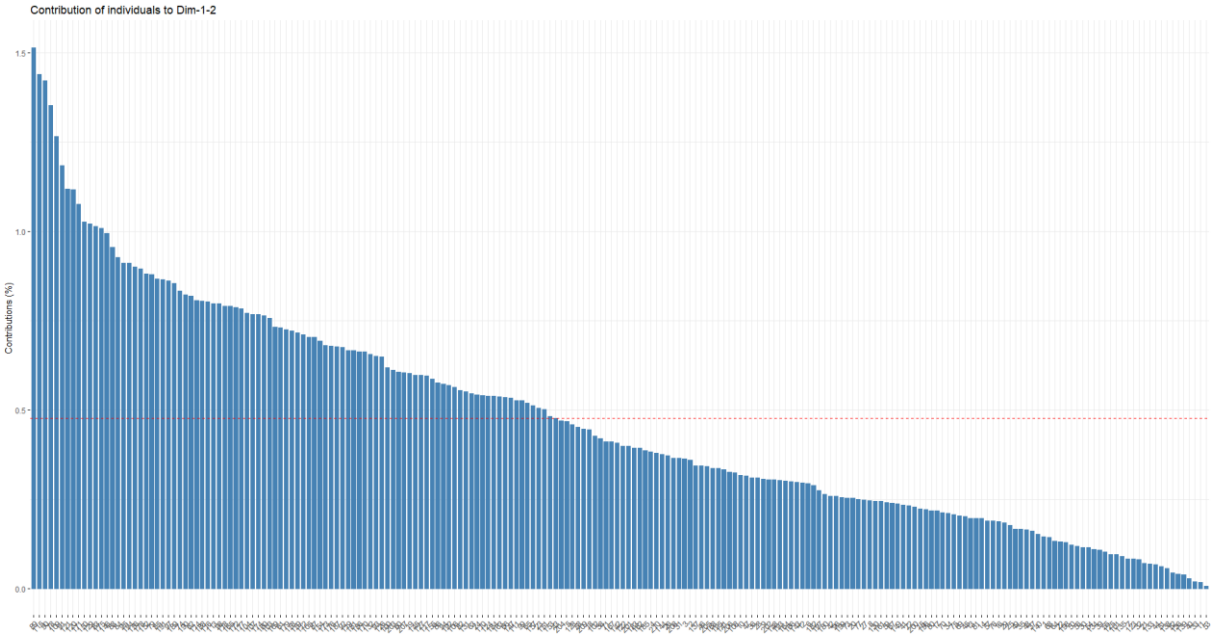
Üçüncü ve dördüncü bileşen için aynı analizi tekrarlarsak değişkenler grafiği;



Üçüncü ve Dördüncü bileşen üzerinde değişkenler grafiği PC3-PC4

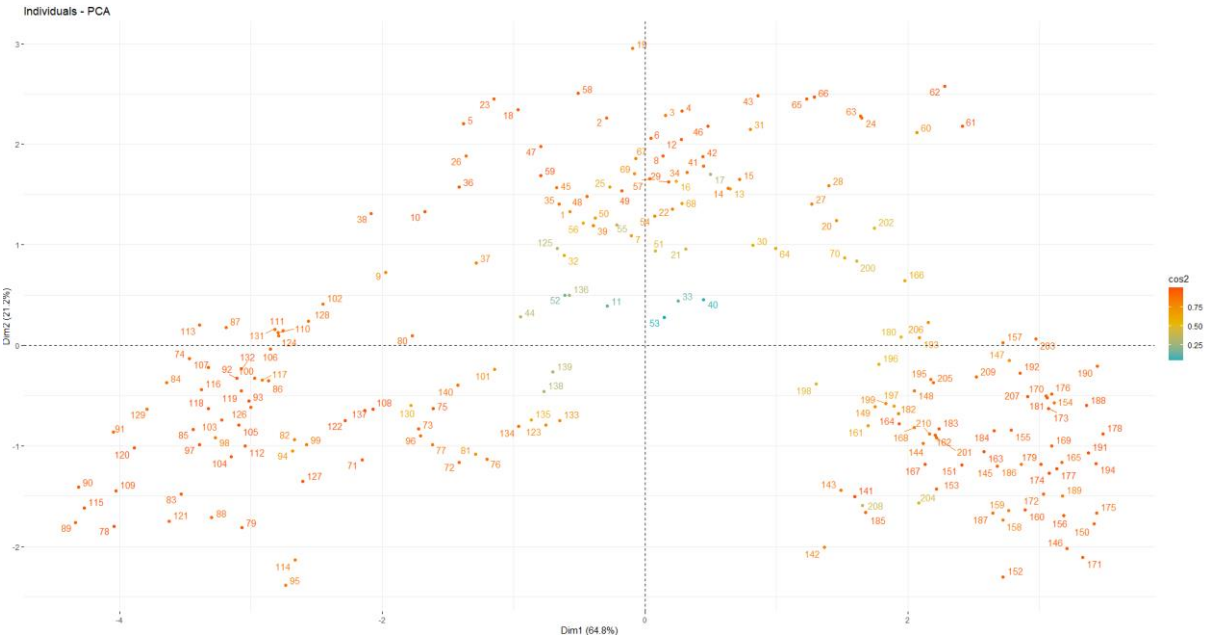
Bu grafikte analizin üçüncü ve dördüncü bileşenlerinde değişkenlerin dağılımı görülmektedir. ***"Compactness"*** ve ***"Asymetry_coef"*** değişkenleri Dim.3 ve Dim.4 üzerinde daha yüksek temil kalitesine sahip olup diğer değişkenlerden farklı bir yönde bulunmaktadır. ***"Class"*** ise Dim.4 üzerinde daha belirgin bir etkiye sahiptir.

PCA analizinde her bir gözlemin (analizde 210 adet) ilk iki bileşene katkı oranlarını ***"fviz_contrib(data.pca, choice = "ind", axes=1:2)"*** komutuyla görselleştirebilmekteyiz. Bu sayede hangi gözlemlerin bileşenler üzerinde daha baskın olduğunu anlaşılabilmektedir.



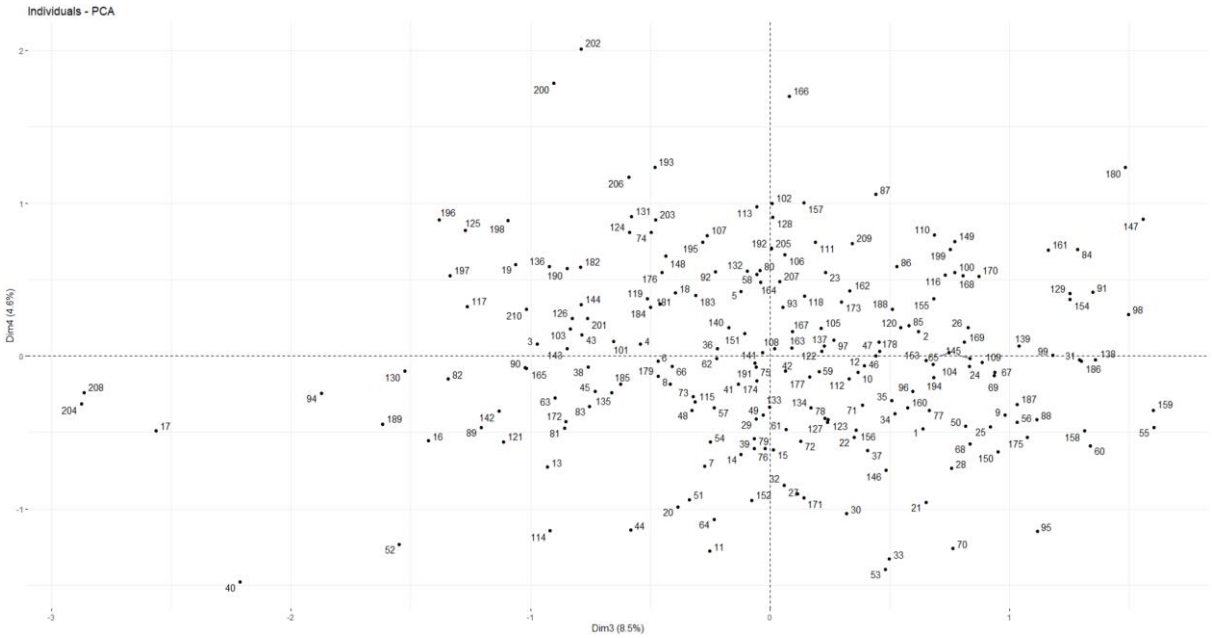
Gözlemlerin bileşenler üzerinde katkı oranları PC1-PC2

Temel bileşen analizinde, gözlemlerin iki boyutlu bir düzlemde dağılımını görselleştirmek için ***"fviz_pca_ind"*** fonksiyonu kullanılmaktadır. Bu fonksiyon ile ilk iki bileşen (PC1 ve PC2) üzerinde benzerliklerin anlaşılmasını sağlamaktadır. Ayrıca her bir gözlem ***"cos2"*** değerine göre renklendirilmektedir.



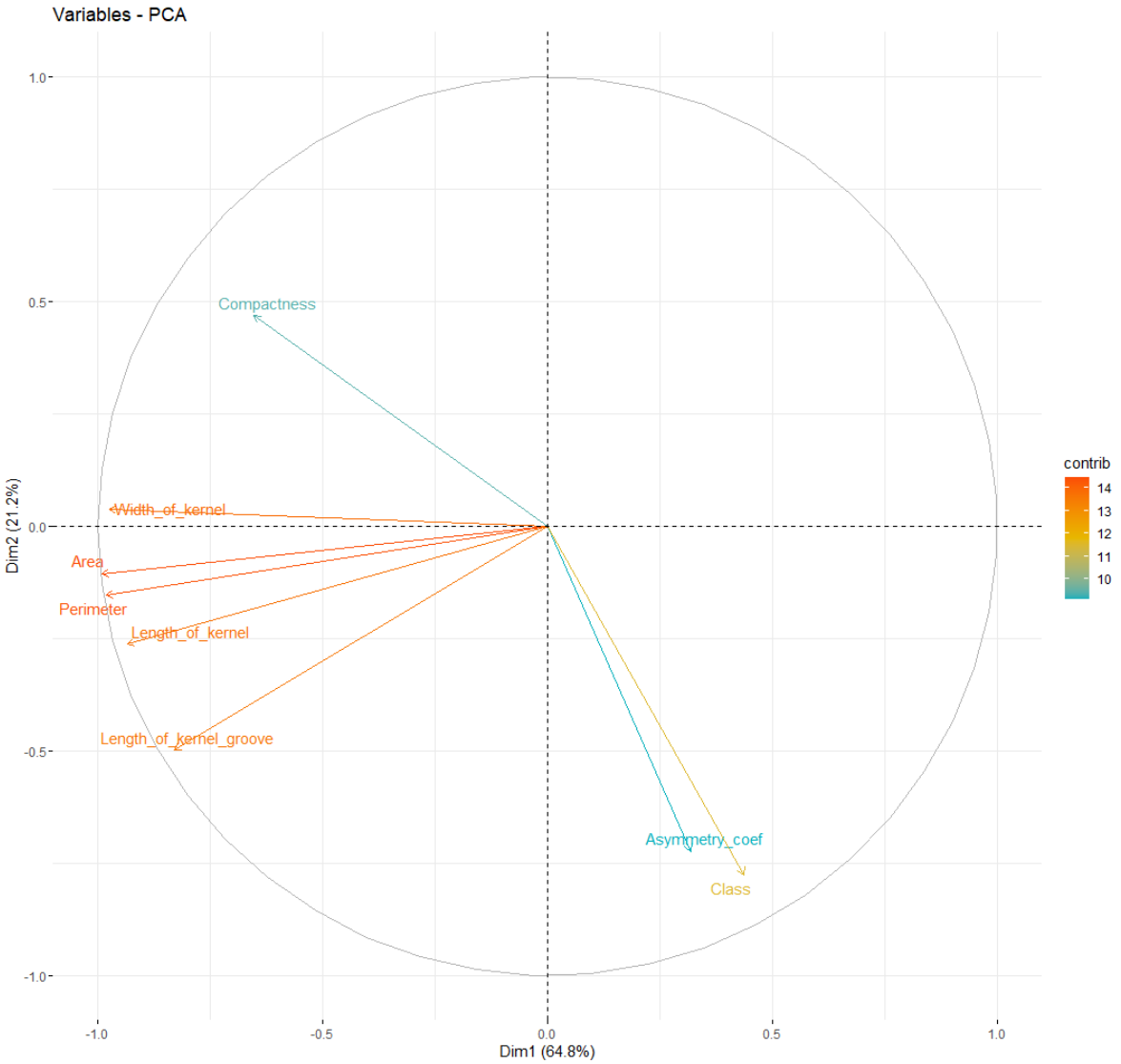
Gözlemler arası benzerlik dereceleri PC1-PC2

Benzerlik grafiğinde ilk iki temel bileşen üzerinde yoğunlaşılmıştır. Gözlemler arasındaki mesafeler benzerlik derecelerini temsil etmekte, yakın gözlemler birbirine yakın gözlemleri ifade etmektedir. Renk skalası, gözlemlerin temsil kalitesini "cos2" değerini yansıtmaktadır. Koyu değerler iyi temsil edilmektedir. Aynı işlemler PC3-PC4 için yapılırsa aşağıdaki grafik elde edilmektedir.



Gözlemler arası benzerlik dereceleri PC3-PC4

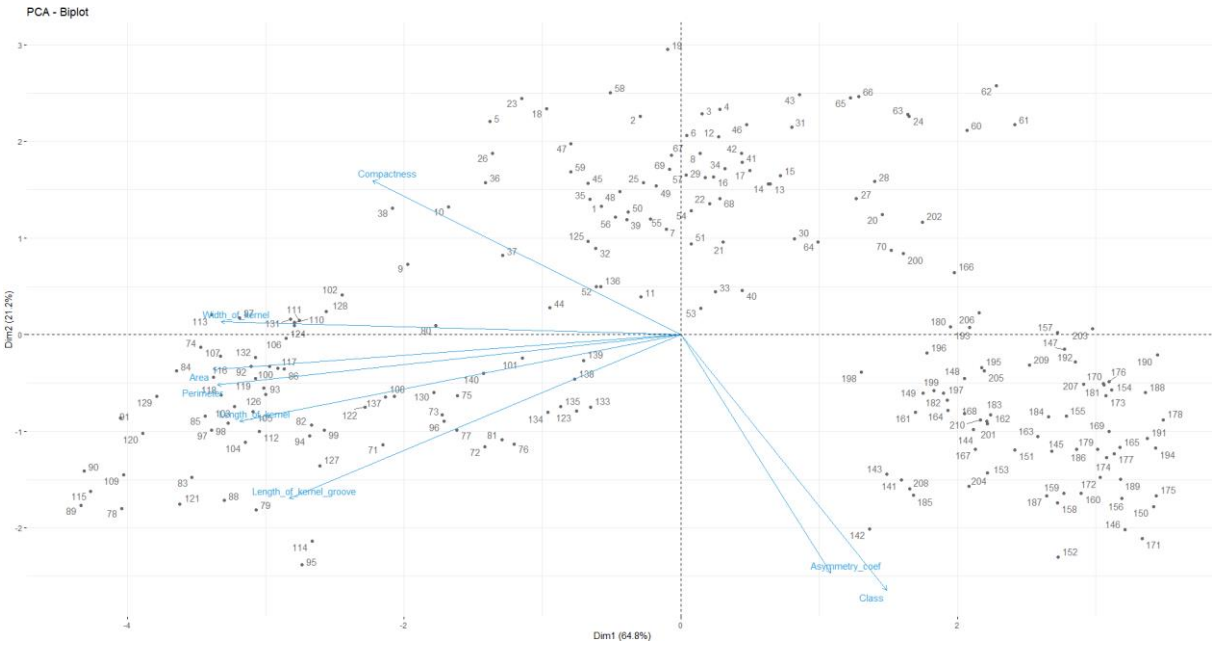
Veri setindeki ilişkileri ve bileşenler üzerindeki katkı oranlarını anlamak için "fviz_pca_var" fonksiyonu ile değişkenlerin düzlemde nasıl konumlandığı ve aralarındaki korelasyon gösterilmektedir. Bu grafikte aynı yöne işaret eden değişkenler pozitif korelasyon, zıt yöne işaret edenler negatif korelasyona sahiptir.



Değişkenlierin ilk iki bileşen üzerindeki dağılım ve katkı

Bu grafikte PCA'da her bir değişkenin ilk iki bileşene katkısı görülmektedir. Negatif ve pozitif korelasyona sahip değişkenler görülebilmektedir. Renk dağılımları değişkenlerin bileşenlerin katkı oranlarını vurgulamaktadır.

Gözlemler ve değişkenleri aynı grafikte göstermek için "fviz_pca_biplot" fonksiyonu ile gözlemler ve değişkenleri aynı düzlemde görselleştirerek biplot oluşturulmaktadır. Gözlemlerin değişkenlerle ilişkisi ve veri yapısı anlaşılabilmektedir.



İlk iki bileşen üzerinde Biplot grafiği

Bu biplot grafiğinde, PCA analizinde gözlemler ve değişkenleri aynı düzlemde gösterilmektedir. Oklar değişkenleri temsil etmektedir. Gözlemler ise değişkenlere göre konumlanmaktadır benzer olanlar yakın konumlanmaktadır.

4. Sonuç

Seeds veri seti üzerinde yapılan analizlerde veri setinin temel özellikleri incelenmiş ve Temel Bileşen Analizi (PCA) ile boyut indirgeme işleminde kullanılabilecek bileşenler belirlenmiştir. İlk olarak veri setinin özet istatistikleri ve standart sapmaları analiz edilerek veri yapısı hakkında genel bilgiler elde edilmiştir. PCA sürecinde veri setinin korelasyon matrisi üzerinden özdeğerler ve özvektörler hesaplanarak temel bileşenler belirlenmiştir. İlk üç temel bileşenin toplam varyansın büyük bir kısmını açıkladığı gözlemlenmiştir. Bu bileşenlerin görselleştirilmesiyle veri setindeki gözlemlerin değişkenler üzerindeki dağılımı ve değişkenlerin veri üzerindeki etkileri incelenmiştir. Çeşitli grafiklerle görselleştirmeler yapılmıştır.

Yapılan PCA analizi, Seeds veri setinde boyut indirgeme için etkili bir yöntem olarak görülmüştür. İlk üç bileşen veri setindeki varyansın büyük bir kısmını (%93) taşıdığı için boyut indirgemede kullanılabilmektedir. Compactness, Asymetry_coef gibi değişkenlerin bileşenler üzerindeki etkisinin yüksek olması veri setinde bul değişkenlerin önemli rol oynadığını göstermektedir. Değişkenlerin farklı yönlerde dağılması korelasyon yapısını ortaya koymaktadır. Bu analiz, veri setindeki ilişkileri daha iyi anlamak ve görselleştirmek için faydalı olduğu değerlendirilmektedir. İlerleyen çalışmalarda kümeleme algoritmaları ile gözlemler arasındaki daha detaylı ilişkiler incelenebilir.

5. Kaynakça

[1] IBM, (2023, 8 Kasım). *Principal Component Analysis (PCA)*. IBM., Erişim Tarihi: 6.11.2024, https://www.ibm.com/topics/principal-component-analysis

[2] Ulutagay, G. (2024). *Denetimsiz Makine Öğrenmesi Yöntemleri : Ders 2 Temel Bileşen Analizi* (Ders Notları). Ege Üniversitesi.

[3] Erzurumlu,E. (2023) *PCA*. Medium., Erişim Tarihi: 6.11.2024, https://medium.com/@erdemerzurumlu/pca-473fe69a2680

[4] Köksoy, O. (2024). *Mathematical Statics Lecture Notes.* Ege Üniversitesi

[5] *Correlation vs. Covariance : What' the Difference?.* Erişim Tarihi: 6.11.2024, https://www.projectpro.io/article/correlation-vs-covariance/489

[6] Margalit, D., Rabinoff, J. & Rolen, L. *Eigenvalues and Eigenvectors.* Georgia Institute of Technology. Erişim Tarihi: 6.11.2024, https://textbooks.math.gatech.edu/ila/eigenvectors.html

[7] UCI Machine Learning Repository. *Seeds Dataset.* Erişim Tarihi: 6.11.2024, https://archive.ics.uci.edu/dataset/236/seeds

6. Ekler

Ek-1

```
# Paketlerin toplu yüklenmesi
> install.packages(c("cluster",     # Kümeleme algoritmaları ve analizleri için
+                    "devtools",    # R'de geliştirme araçları için
+                    "pastecs",     # İstatistiksel özetler ve tanımlayıcı analizler
için
+                    "corrplot",    # Korelasyon matrislerinin görselleştirilmesi için
+                    "factoextra"   # PCA ve diğer çok değişkenli analizlerin
görselleştirilmesi için
+ ))
# Yüklü paketlerin çağrılması
library(cluster)     # Kümeleme analizleri (ör. K-means)
library(devtools)    # R için geliştirme araçları
library(pastecs)     # Betimleyici istatistikler ve veri özetleri
library(corrplot)    # Korelasyon matrisi görselleştirmesi
library(factoextra)  # PCA sonuçlarının görselleştirilmesi
 # TXT dosyasını yükleme (herhangi bir sınırlayıcıyla)
> data <- read.table("D:/seeds_pca/seeds/seeds_dataset.txt",
+                    header = FALSE, sep = "", stringsAsFactors = FALSE)
>
> # Kolon isimlerini belirleme (sınıf sütununu çıkarmak için)
> colnames(data) <- c("Area", "Perimeter", "Compactness",
+                     "Length_of_kernel", "Width_of_kernel",
+                     "Asymmetry_coef", "Length_of_kernel_groove", "Class")
>
#boyut ve ilk 6 satır
dim(data)
head(data)
#Veri setinin temel istatistiklerii, sütındaki standart sapma değerleri, kutu grafiği
vb.)
summary(data)
apply(data,2,sd)
boxplot(data, horizontal=TRUE)
stat.desc(data)
```

```
require(graphics)

pairs(data)

#Veri setinin temel istatistiklerii, sütındaki standart sapma değerleri, kutu grafiği
vb.)

#manual

corr <- cor((data), method = "pearson")

corr

corrplot.mixed(corr, lower="pie",upper="number")

attach(data)

boxplot(data)

data.cor <- cor(data)

data.eigen <- eigen(data.cor)

data.eigen

eigenvalues <- data.eigen$value

prop.var <- data.eigen$value/sum(data.eigen$values )

cum.prop.var <- cumsum(prop.var)

rbind(eigenvalues,prop.var,cum.prop.var)

plot(prop.var , xlab=" Principal Component ", ylab=" Proportion of

Variance Explained ", ylim=c(0,1) ,type='b')

lines(cum.prop.var, type='b', col="red")

#manual

##paket üstünden hesaplama

data.pca <- prcomp(data, center = TRUE, scale. = TRUE)  ##korelasyon matrisi üzerinden

summary(data.pca)

(data.pca$sdev)^2 ##özdeğerler

biplot(data.pca)

cor(data)

data.pca$rotation  ##özvektörleri verir.

data.pca$center  ##değişkenlerin ortalamaları

data.pca$scale   ##değişkenlerin standart sapmaları

##paket üstünden hesaplama

##görselleştirme

fviz_eig(data.pca, addlabels = TRUE)

# Eigenvalues

eig.val <- get_eigenvalue(data.pca)

eig.val
```

```
# Results for Variables
res.var <- get_pca_var(data.pca)
res.var$coord          # Coordinates
res.var$contrib        # Contributions to the PCs
res.var$cos2           # Quality of representation
library("corrplot")
corrplot(res.var$cos2, is.corr=FALSE) # Quality of representation
##Variable contributions to the principal axes:
  # Contributions of variables to PC1
fviz_contrib(data.pca, choice = "var", axes = 1, top = 4)
# Contributions of variables to PC2
fviz_contrib(data.pca, choice = "var", axes = 2, top = 4)
fviz_pca_var(data.pca, col.var = "cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE # Avoid text overlapping
)
##PC3-PC4
fviz_pca_var(data.pca, axes = c(3, 4), col.var = "cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE # Avoid text overlapping
# Results for individuals
res.ind <- get_pca_ind(data.pca)
res.ind$coord [,1:2]          # Coordinates
res.ind$contrib        # Contributions to the PCs
res.ind$cos2           # Quality of representation
fviz_contrib(data.pca, choice = "ind", axes = 1:2)
fviz_pca_ind(data.pca,
             col.ind = "cos2", # Color by the quality of representation
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE     # Avoid text overlapping
)
##PC3-PC4
fviz_pca_ind(data.pca, axes = c(3, 4), col.var = "cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE # Avoid text overlapping
)
```

```
##Graph of variables. Positive correlated variables point to the same side of the
plot.

##Negative correlated variables point to opposite sides of the graph.

fviz_pca_var(data.pca,

             col.var = "contrib", # Color by contributions to the PC

             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),

             repel = TRUE     # Avoid text overlapping

)

##Biplot of individuals and variables

fviz_pca_biplot(data.pca, repel = TRUE,

                col.var = "#2E9FDF", # Variables color

                col.ind = "#696969"  # Individuals color

)

##görselleştirme
```