T.C

DOKUZ EYLÜL ÜNİVERSİTESİ

FEN FAKÜLTESİ

İST 4138 Statistical Methods in Data Mining

2017285019 ALPER ENGİN

2017285037 ATADENİZ SAYAR

2017285023 CEM GÖRENER

2017285066 ÇAĞATAY GÜLMEZ

2018285023 YUNUS ERGÜN

Veri Seti :

Bağımsız Değişkenler :

Alan : Sürekli

Çevre : Sürekli

Yoğunluk : Sürekli

Çekirdek Uzunluğu : Sürekli

Çekirdek Genişliği : Sürekli

Asimetri Katsayısı : Sürekli

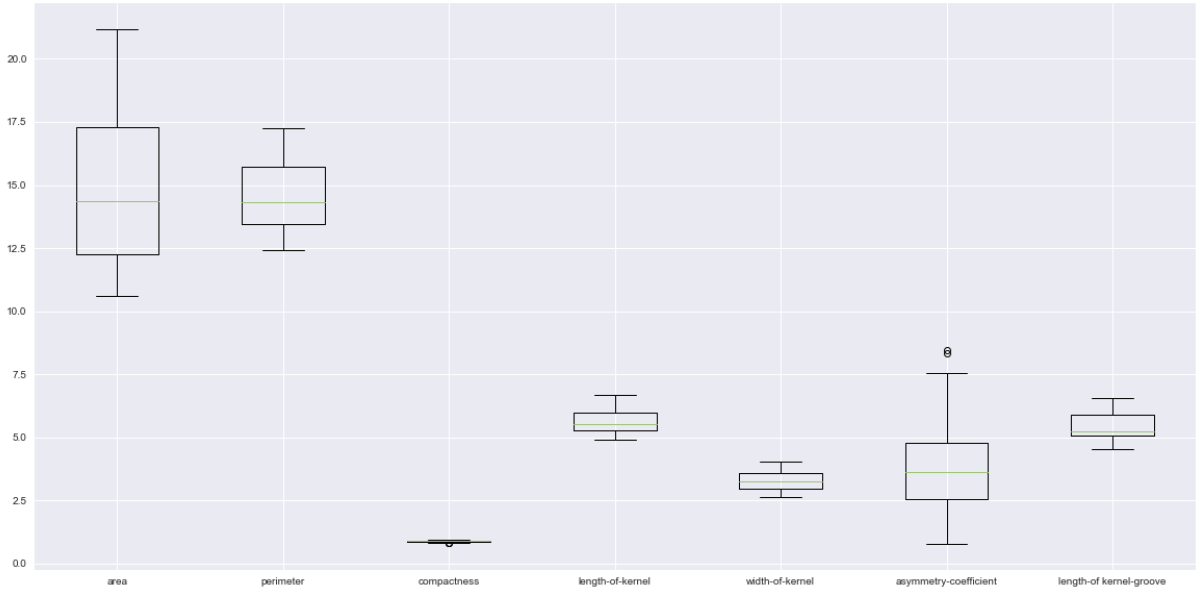
Çekirdek Oluğunun Uzunluğu : Sürekli

Bağımlı Değişken :

Tohum Tipi : `"Kama" = 0,"Rosa" = 1,"Canadian" = 2`

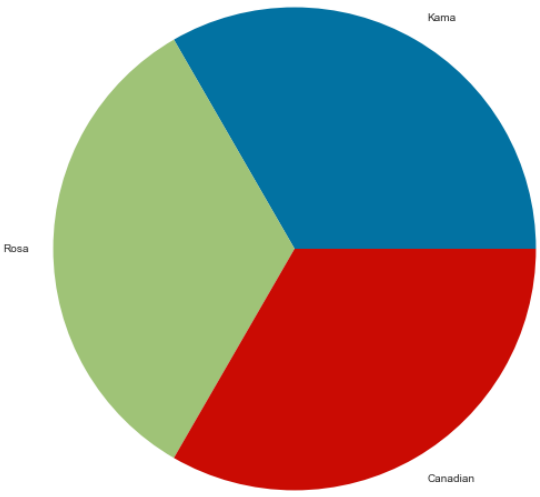
Sayısal Değişkenlerin Tanımlayıcı İstatistikleri ve Kutu Grafiği

	area	perimeter	compactness	length-of-kernel	width-of-kernel	asymmetry-coefficient	length-of kernel-groove
count	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000
mean	14.847524	14.559286	0.870999	5.628533	3.258605	3.700201	5.408071
std	2.909699	1.305959	0.023629	0.443063	0.377714	1.503557	0.491480
min	10.590000	12.410000	0.808100	4.899000	2.630000	0.765100	4.519000
25%	12.270000	13.450000	0.856900	5.262250	2.944000	2.561500	5.045000
50%	14.355000	14.320000	0.873450	5.523500	3.237000	3.599000	5.223000
75%	17.305000	15.715000	0.887775	5.979750	3.561750	4.768750	5.877000
max	21.180000	17.250000	0.918300	6.675000	4.033000	8.456000	6.550000



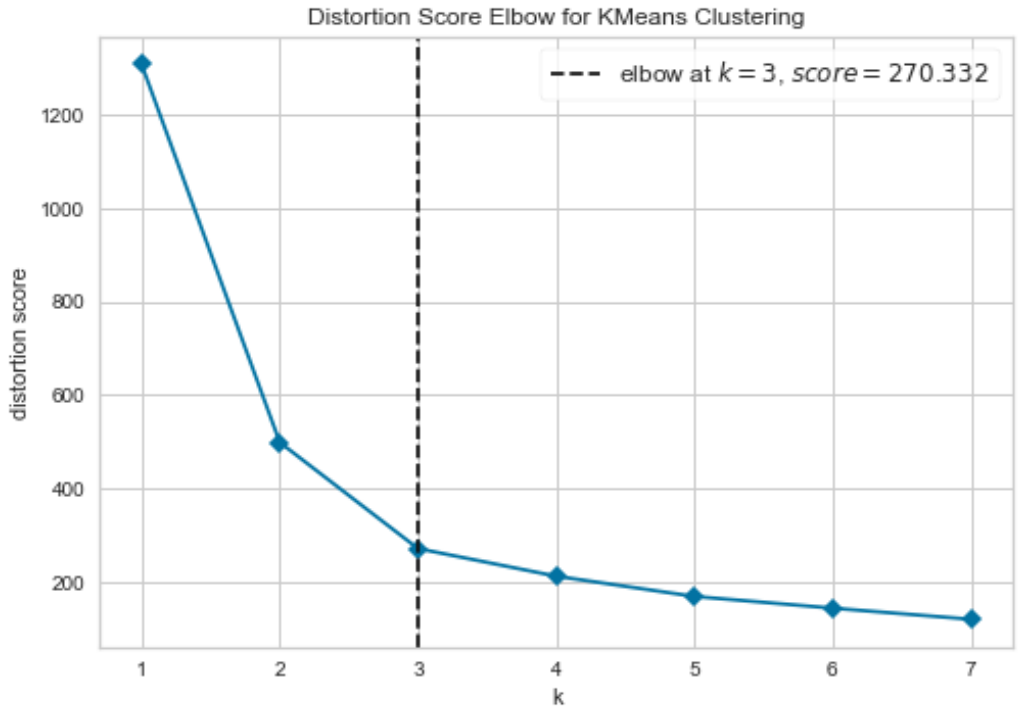
Korelasyon Grafiği

Bağımlı Değişkenimiz Olan Tohum Tipinin Pasta Grafiği

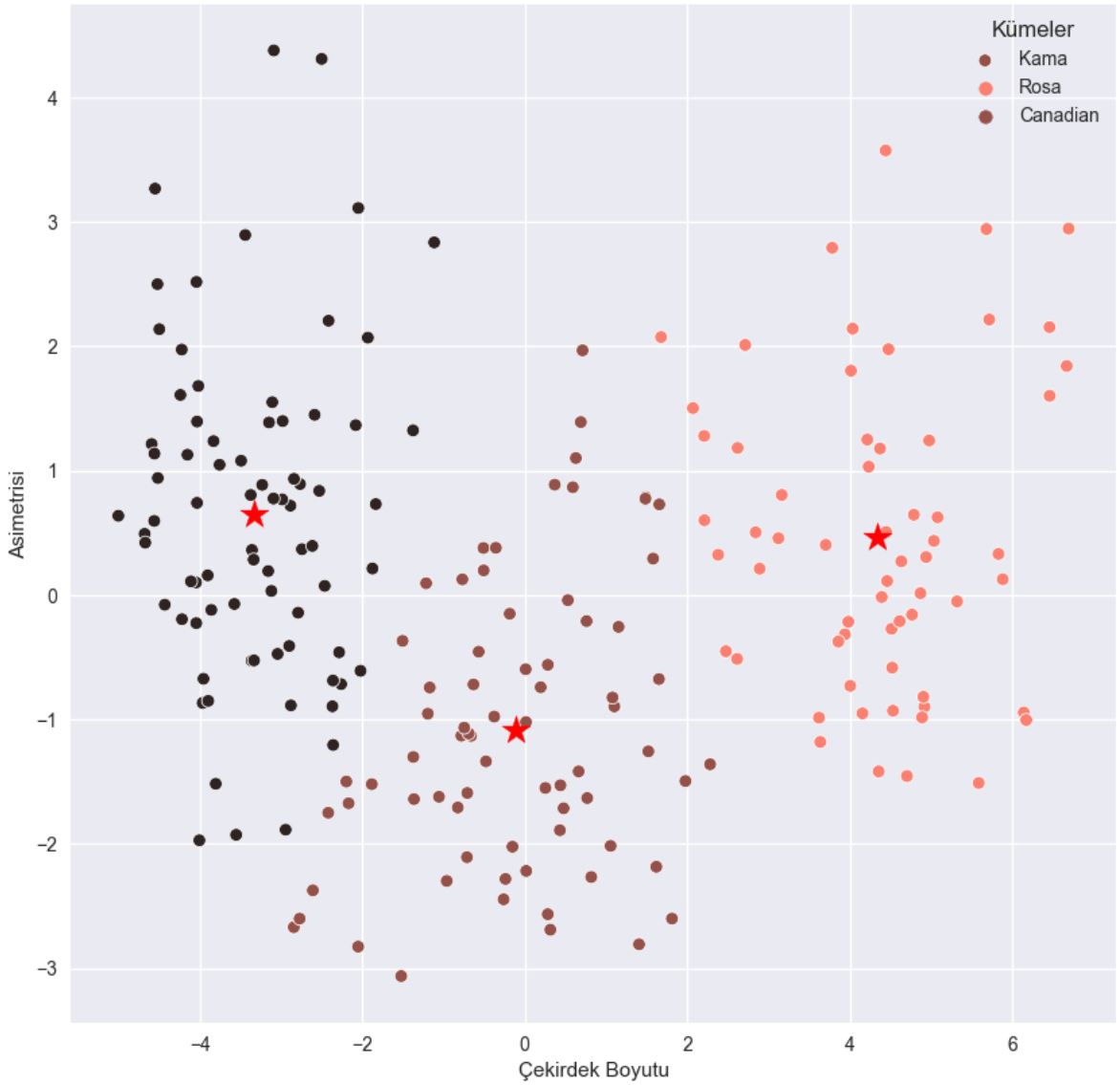


Temel Bileşenler Analizi İle Veri Setini 2 Boyuta Düşürme

	Variables	Çekirdek Boyutu	Asimetrisi
0	area	0.884229	0.100806
1	perimeter	0.395405	0.056490
2	compactness	0.004311	-0.002895
3	length-of-kernel	0.128544	0.030622
4	width-of-kernel	0.111059	0.002372
5	asymmetry-coefficient	-0.127616	0.989410
6	length-of kernel-groove	0.128966	0.082233



Elbow grafiğinin sonuçlarına bakıldığında K-Means kümeleme yöntemi için en uygun küme sayısı 3 olarak bulunmuştur.



K-Means Skorları

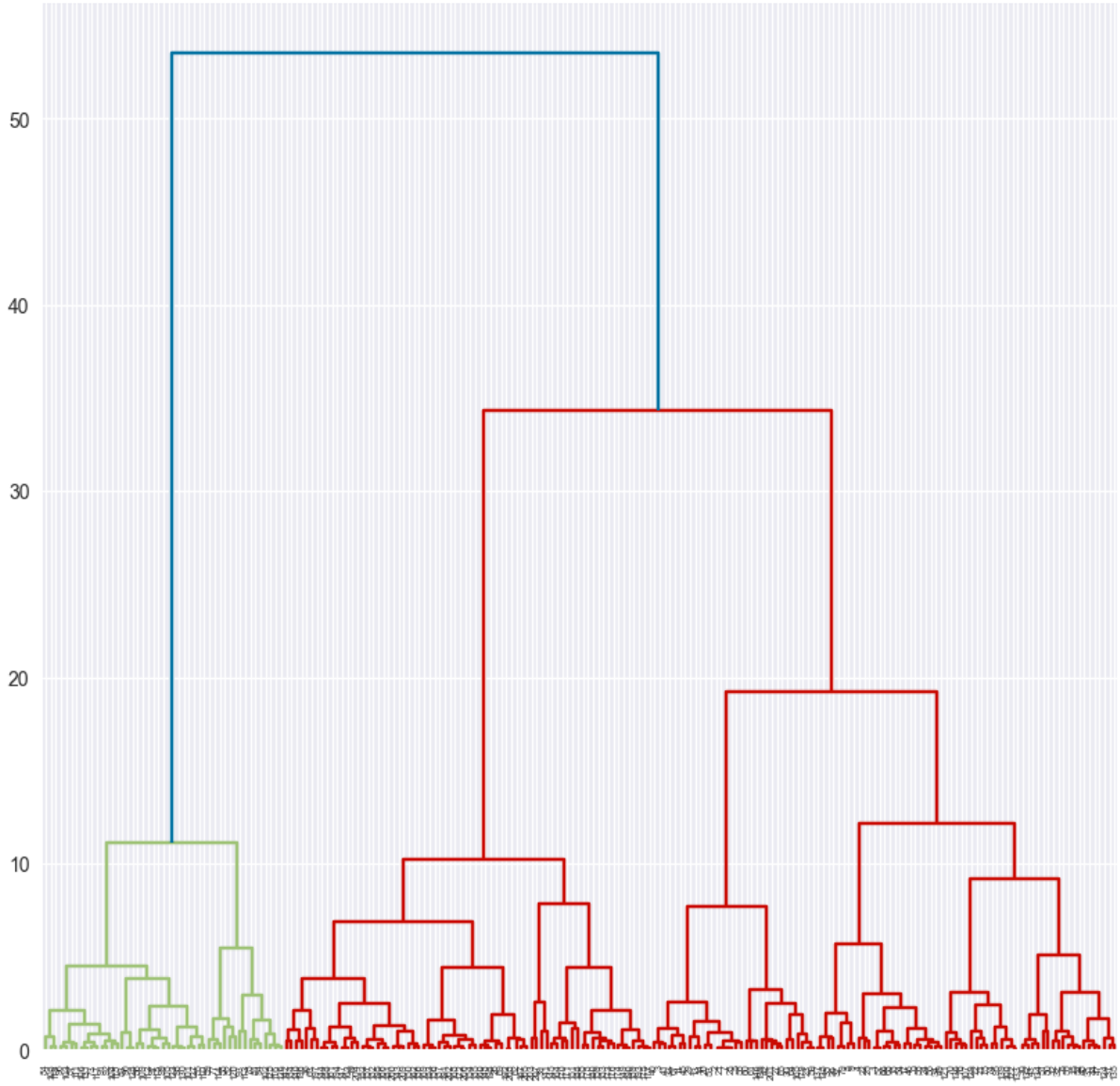
Silhouette : 0.4802142699427175

Completeness : 0.696395547296022

Homogeneity : 0.6934607041029826

V_Ölçüsü : 0.696395547296022

Hiyerarşik Kümeleme



Hiyerarşik kümeleme yönteminin dendogramına bakıldığında en uygun küme sayısının 3 olacağına karar verilmiştir.

“Linkage = ward” Hiyerarşik Kümele Skorları

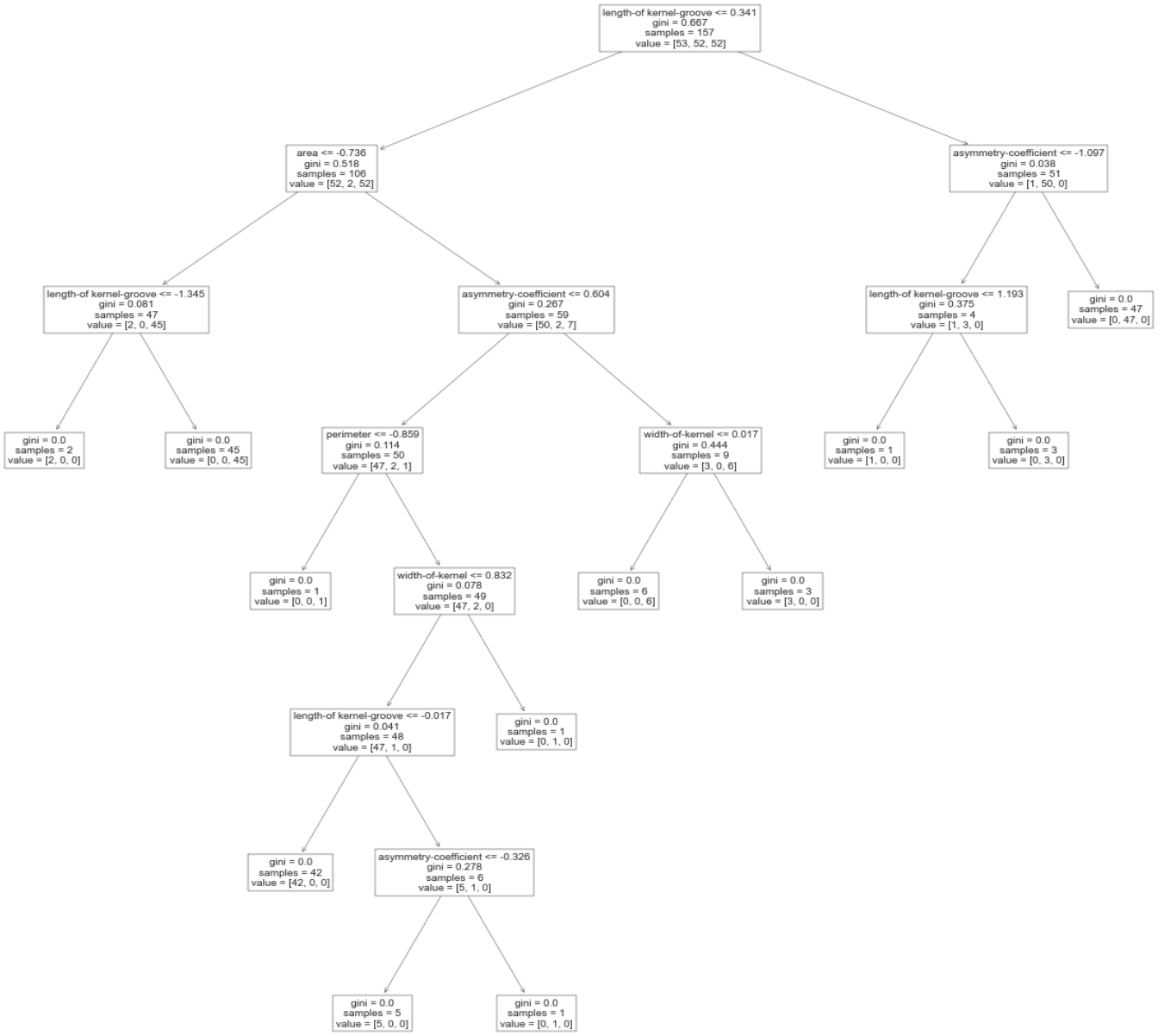
Silhouette : 0.43306620092682413

Completeness : 0.6193332468707496

Homogeneity : 0.6000556739749708

V_Ölçüsü : 0.6193332468707496

Karar Ağacı Algoritması

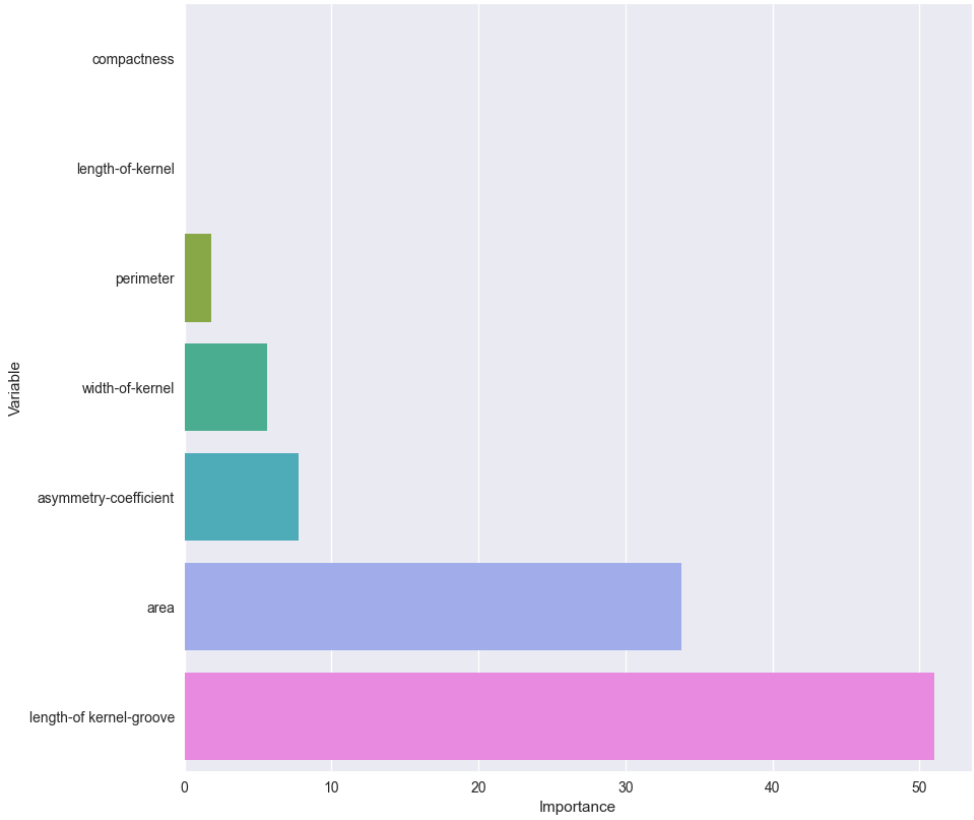


	precision	recall	f1-score	support
0	0.88	0.88	0.88	17
1	0.94	0.94	0.94	18
2	0.94	0.94	0.94	18
accuracy			0.92	53
macro avg	0.92	0.92	0.92	53
weighted avg	0.92	0.92	0.92	53

Metrikler yakından incelendiğinde TPR değerlerine bakıldığında her sınıf için %80 üzerinde olması ve doğruluk oranının %92 çıkması ile beraber modelin oldukça yeterli olduğunu ve başarılı bir şekilde sınıflama yapabildiğini söyleyebiliriz.

En iyi parametrelerin bulunması için GridSearchCV algortiması ile k = 10 olacak şekilde k-fold crossvalidation yöntemi ile eğitim verisi üzerinde çalışılmış.
En olarak seçilen parametreler ile aşağıdaki model kurulmuştur.

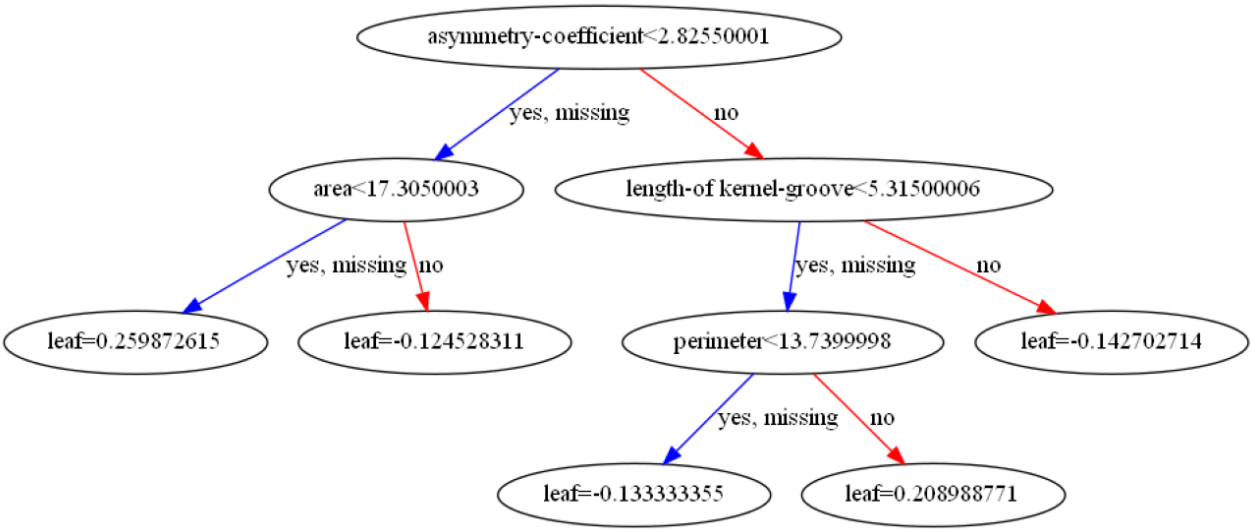
RandomForestClassifier(ccp_alpha=0.01,criterion="gini",max_features="sqrt",max_samples=6,n_estimators=50) parametreleri ile kurulan modelimizin sınıflama metrikleri aşağıdaki gibidir.
Rassal ormanlar modeli için değişkenlerin önem düzeyi yani ağaçtaki köke yakın olma durumu grafikte gözükmektedir.



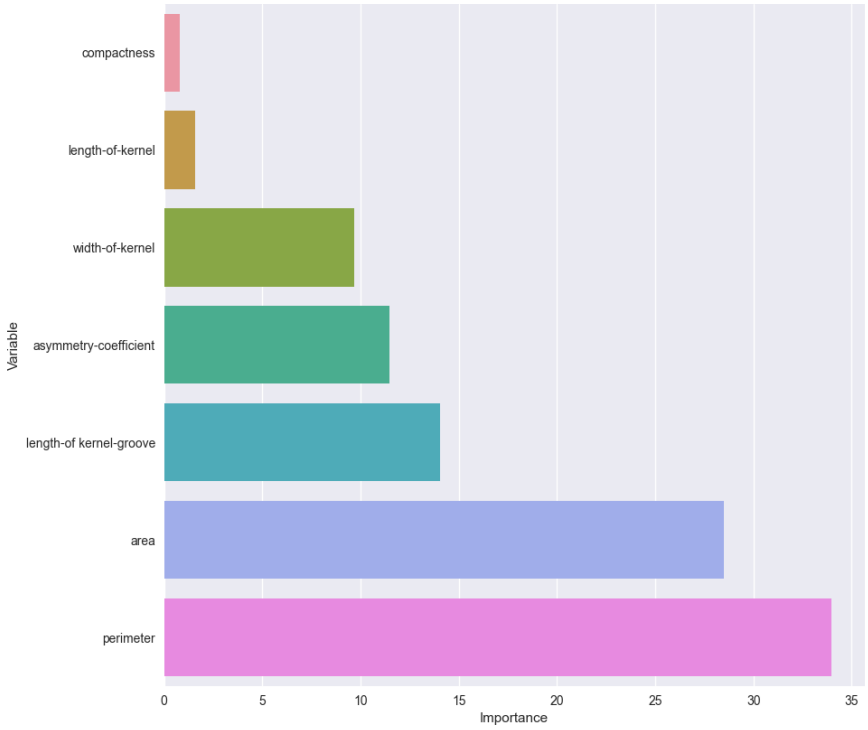
	precision	recall	f1-score	support
0	0.85	0.65	0.73	17
1	0.89	0.94	0.92	18
2	0.81	0.94	0.87	18
accuracy			0.85	53
macro avg	0.85	0.85	0.84	53
weighted avg	0.85	0.85	0.84	53

TPR değerlerine bakıldığında “Kama” çekirdek tipi için düşük diğer çekirdek tipleri için oldukça iyi olduğunu, doğruluk oranı olarak ise %85 doğrulukla çekirdekleri birbirinden ayırt edebildiğini söyleyebiliriz.

XGBoost Modeli



Grafiğe bakıldığında XGBoost modeli için değişkenlerin önem düzeyi yani ağaçtaki köke uzak olma durumu grafikte gözükmektedir.



	precision	recall	f1-score	support
0	0.82	0.82	0.82	17
1	0.89	0.89	0.89	18
2	0.94	0.94	0.94	18
accuracy			0.89	53
macro avg	0.89	0.89	0.89	53
weighted avg	0.89	0.89	0.89	53

Metrikler yakından incelendiğinde TPR değerlerine bakıldığında her sınıf için %80 üzerinde olması ve doğruluk oranının %89 çıkması ile beraber modelin oldukça yeterli olduğunu ve başarılı bir şekilde sınıflama yapabildiğini söyleyebiliriz.

SONUÇ

Kullandığımız veri seti nezdinde konuşmak gerekirse, kullanmış olduğumuz kümeleme yöntemleri arasından en etkili ve başarılı kümele yapan yöntem K-Means olmuştur.

Kullanmış olduğumuz sınıflandırma algoritmaları arasında XGBoost ve Karar Ağaçları algoritmaları çok iyi sonuçlar çıkarmakla birlikte XGBoost'un karmaşık veri yapıları uğraşma gücü nispeten ufak bir veri setinde karar ağaçları algoritmasına kıyasla daha zayıf sonuçlar doğurmasına yola açmıştır. Sonuç olarak sınıflandırma modelleri arasında modelimiz için en iyi yöntemin karar ağaçları olduğunu söyleyebiliriz.