BİR TARIM ÜRÜNÜ SINIFLANDIRMA PROBLEMİNE UYGULANAN DESTEK VEKTÖR MAKİNESİ MODELİNDE ÇEKİRDEK FONKSİYONLARININ ETKİLİLİĞİNİN ARAŞTIRILMASI

Ezo Erdem

22 Haziran 2023

İçindekiler

1 GİRİŞ

1.1 ÖZET

Bu tez kapsamında Destek Vektör Makinelerinin (Support Vector Machine) çekirdek fonksiyonları (Kernel Function) incelenmiştir. Destek Vektör Makineleri sınıflandırma problemlerinde kullanılan bir yöntemdir. Algoritmanın çalışması esnasında verilerin türüne bağlı olarak çekirdek fonksiyonlar da kullanılabilmektedir. Eğer sınıflandırma işleminde, tam ayrıştırılabilir veriler kullanılırsa genellikle tüm veriler bir hiper düzlem ile sınıflandırılabilmektedir. Fakat, eğer tam ayrıştırılamayan veriler kullanılırsa, çoğunlukla aynı boyutta tek bir düzlem ile sınıflandırılamamaktadır. Bu nedenle de farklı çekirdek fonksiyonları kullanılmaktadır. Destek Vektör Makinelerinin literatürde bu kadar başarılı olmalarının sebebi çekirdek yöntemlerinin kullanılmasıdır.
Sınıflandırma problemlerinin çözümü için geliştirilen makine öğrenimi algoritmasının seçiminde dikkat edilecek önemli unsurlardan biri, algoritmanın genelleme performansıdır. Genelleme performansı, eğitim verisi, bağımsız niteliklerin sayısı/yapısı, model seçimi ve parametre seçimi gibi faktörlere bağlıdır.
Destek Vektör Makineleri, birçok sınıflandırma probleminin çözümünde başarıyla uygulanmış ve genelleme performansı yüksek ve etkin makine öğrenimi algoritmalarından biri olarak literatürdeki yerini almıştır.
Destek Vektör Makineleri'nin en önemli avantajı, sınıflandırma problemini bir optimizasyon problemine dönüştürüp çözmesidir. Böylece problemin çözümüne ilişkin öğrenme aşamasında işlem sayısı azalmakta ve diğer teknik/algoritmalara göre daha hızlı çözüme ulaşılmaktadır (Osowski, Siwekand ve Markiewicz, 2004). Teknik bu özelliğinden dolayı, özellikle büyük hacimli veri setlerinde büyük avantaj sağlamaktadır. Ayrıca optimizasyon temelli olduğundan sınıflandırma performansı, hesaplama karmaşıklığı ve kullanışlılık açısından diğer tekniklere göre daha başarılıdır (Nitze, Schulthess ve Asche, 2012).
Çeşitli veri setleri için sınıflandırma probleminin çözümüne ilişkin Destek Vektör Makineleri'nin uygulanma aşamasında çekirdek fonksiyonu seçimi çok önemlidir.

1.2 BU ALANDA YAPILMIŞ ÇALIŞMALAR

Sınıflandırma problemlerinin Destek Vektör Makineleriyle (DVM) çözümü ve en iyi çekirdek fonksiyonunun seçimi için, literatürde en sık kullanılan 11 adet veri seti Şekil 1 de gösterilmiştir. Bu veri setleri UCI (Machine Learning Repository) makine öğrenimi veri tabanı sisteminden bulunmuştur. Söz konusu veri setleri farklı sınıflandırma algoritmalarının uygulanması ve performanslarının kar-

şılaştırılmasında bir kriter olarak kabul edilmektedir (Huang ve Wang, 2006).
Sonuç olarak literatürde en sık kullanılan veri setleri; dördü bankacılık, üçü tıp, bir adet bilgisayar sistemleri, fizik, kimya, biyoloji ve hukuk olmak üzere yedi farklı alandaki sınıflandırma problemlerine ilişkindir. Bu çalışmada bu klasik veri setlerinden farklı bir veri seti olan "Dry Bean" veri seti kullanılmıştır.
Ayrıca literatürde birçok çekirdek fonksiyonu tanımlanmıştır. Fakat her çekirdek fonksiyonu Destek Vektör Makinelerinde kullanımı uygun olmamaktadır. Dolayısıyla, DVM'nin uygulanmasında çekirdek fonksiyonlarının seçilmesi kritik bir rol oynamaktadır. Bu çalışmada, DVM için radyal tabanlı, polinomiyal, lineer ve sigmoid çekirdek fonksiyonları kullanılmaktadır.

No	Veri Seti	Örnek Sayısı	Nitelik Sayısı		Sınıf Sayısı
			Kategorik	*Sayısal*	
1	Australian Credit Approval-Statlog	690	8	6	2
2	Bank	4521	9	7	2
3	German Credit-Statlog	1000	20		2
4	Hearth Disease-Statlog	270	6	7	2
5	Ionosphere	351		34	2
6	Pima Indian Diabets (PIM)	768	-	8	2
7	Spambase (Spam)	4601	-	57	2
8	Wisconsin Breast Cancer-WBC-orijinal	699	-	10	2
9	Glass	214	-	9	6
10	Iris Plant (Iris)	150	-	4	3
11	Wine	178	-	13	3
12	**Türkiye kredi verisi (Türkiye)**	167	13	4	2

Şekil 1: Literatürde En Sık Kullanılan Veri Setleri Ve Özellikleri

Bu çalışmada, DVM algoritmasının uygulanmasında, çekirdek fonksiyonlarının optimal hiper parametre değerlerinin belirlenmesinde grid arama yöntemi kullanılmıştır. Yöntemin uygulanmasında her bir parametre için alt sınır, üst sınır ve belirli bir aralık değeri belirlenir. Parametre değerleri, sınır değerleri içinde belirlenen aralık kadar atlayarak her bir değer noktası için algoritmaya ilişkin bir sınıflandırma performansı belirler. En iyi sınıflandırma performansını veren parametre değerleri optimal hiper parametre değerleri olarak belirlenir. Lineer, radyal tabanlı (C, γ), polinomiyal

(C, γ, α, d) ve sigmoid (C, γ, α) olmak üzere dört çekirdek fonksiyonu için çalışmamızda belirlenen parametre değer aralıkları Şekil 2 de verilmiştir. Söz konusu çekirdek fonksiyonlarına ilişkin C ceza parametresinin alt sınırı $0,0001$ ve üst sınırı 5000 olarak belirlenmiştir. Parametrelere ilişkin artış aralıkları 1 olarak alınmıştır. Ancak bu aralık logaritmik artışı göstermektedir. Örneğin, alt sınır $2^{(-13)}(0,0001)$ değerinden başlarsa bir sonraki parametre değeri $2^{(-12)}(0,0002441406)$ olarak alınmaktadır. Diğer parametre değerleri için benzer şekilde yorumlanabilir.

Parametreler	*Alt Sınır*	*Üst Sınır*	*Aralık*
C	0.0001	5000	$[2^{-13}, 1, 2^{13}]$
γ	0.001	500	$[2^{-9}, 1, 2^{9}]$
α	0.0001	50	$[2^{-13}, 1, 2^{6}]$
d	1	3	1

Şekil 2: Grid Arama İçin Belirlenen Parametre Değer Aralıkları

Veri Setleri	Çekirdek fonksiyonları	Çekirdek değerleri fonksiyonlarına ilişkin parametre				Eğitim (%)	Test (%)	Ortalama (%)
		Ć	γ	Degree	Coef			
Austaralian Credit	Linear	0,1	-	-	-	86,20	86,20	86,2
	Radial	0,1	0,435603	-	-	88,52	86,47	87,495
	Polinomial	0,25763	0,03684	3	0,215444	87,77	87,09	87,43
	Sigmoid	0,39789	0,16681	-	0,02154	86,20	86,17	86,185
German Credit	Linear	0,39789	-	-	-	78,60	76,50	77,55
	Radial	41,01832	0,003327	-	-	79,40	76,90	78,15
	Polinomial	0,199474	0,003327	3	12,91549	78,80	76,70	77,75
	Sigmoid	25,06597	0,011081	-	0,46416	76,90	77,00	76,95
Glass	Linear	50	-	-	-	72,90	61,22	67,06
	Radial	12,56605	15,02665	-	-	95,46	73,37	84,415
	Polinomial	1,34876	4,516	3	4,64159	91,6	72,00	81,8
	Sigmoid	50	0,122583	-	0,21544	64,00	62,6	63,3
Hearth disease	Linear	0,103702	-	-	-	87,40	83,7	85,55
	Radial	0,79370	0,03684	-	-	86,67	85,19	85,93
	Polinomial	25,06595	0,01107	3	0,00464	85,93	84,07	85
	Sigmoid	0,19947	0,12258	-	0,02154	84,81	84,07	84,44

Şekil 3: Çekirdek Fonksiyonlarının DVM Performansları Ve Parametre Değerleri

Pima Indian Diabet	Linear	0,383166	-	-	-	77,99	77,21	77,6
	Radial	1,64183	0,12258	-	-	78,13	77,73	77,93
	Polinomial	0,1	0,011072	3	10	77,99	76,95	77,47
	Sigmoid	25,06595	0,03684	-	1	77,99	77,08	77,535
Ionosphere	Linear	1	-	-	-	87,59	81,15	84,37
	Radial	1	0,1	-	-	94,32	91,30	92,81
	Polinomial	1	0,1	1	0,01	87,23	84,06	85,645
	Sigmoid	1	0,1	-	0,01	80,85	79,71	80,28
Iris Plant	Linear	1502,6652	-	-	-	98	96,7	97,35
	Radial	1,1071732	1,3572	-	-	97,3	97,3	97,3
	Polinomial	3,684032	50	1	0	98	98	98
	sigmoid	40,7886	0,12258	-	0,07743	97,3	97,3	97,3
Spambase	Linear	10	-	-	-	94,60	92,82	93,71
	Radial	10	0,01	-	-	93,45	94,00	93,725
	Polinomial	10	0,01	1	0,01	92,50	92,30	92,4
	Sigmoid	10	0,01	-	0,001	88,33	87,54	87,935
	Linear	0,199473	-	-	-	99,99	96,01	98

Şekil 4: Çekirdek Fonksiyonlarının DVM Performansları Ve Parametre Değerleri

Wine	Radial	0,22901	0,4	-	-	99,99	97,20	98,595
	Polinomial	0,1	1,3572	3	0,0000	100	97,20	98,6
	Sigmoid	12,5660	0,12258	-	0,0000	100	98,3	99,15
Wisconsin	Linear	0,09629	-	-	-	97,22	96,93	97,075
Breast	Radial	0,397897	0,4078	-	-	97,22	97,07	97,145
Cancer	Polinomial	25,06596	0,12258	1	1	97,66	97,07	97,365
(WBC)	Sigmoid	0,1	0,40788	-	0,21544	97,66	97,66	97,66
	Linear	100	-	-	-	89,33	85,50	87,415
	Radial	100	0,01	-	-	89,60	89,85	89,725
Bank	Polinomial	100	0,01	1	0,001	89,33	85,50	87,415
	Sigmoid	100	0,01	-	0,001	85,82	82,60	84,21
	Linear	0,09321	-	-	-	97,01	86,83	91,92
Türkiye	Radial	0,806234	0,15590	-	-	100	89,82	94,91
Kredi verisi	Polinomial	0,06787	1,35721	1	0,0000	97,01	86,23	91,62
	Sigmoid	3,684	0,03684	-	4,64159	91,02	89,82	90,42

Şekil 5: Çekirdek Fonksiyonlarının DVM Performansları Ve Parametre Değerleri

Literatürdeki çalışmalarda, test verileri için çekirdek fonksiyonlarının sınıflandırma performansları incelendiğinde (Şekil 3-4-5), glass (%73,37), hearth disease (%85,19), Pima Indian diabet (%77, 73), ionosphere (%91,30), spambase (%94,00), bank (%89,85) ve Türkiye kredi (%89,82) veri setlerinin performanslarına bakarak radyal tabanlı çekirdek fonksiyonunun en iyi performansa sahip çekirdek fonksiyonu olduğu gözlemlenmiştir. Polinomiyal çekirdek fonksiyonu için Australian credit (%87,09) ve Iris (%98) veri setinin daha başarılı sonuçlar verdiği görülmektedir. German credit (%77), wine (%98,3) ve WBC (%97,66) veri setleri için sigmoid çekirdek fonksiyonunun daha başarılı sınıflandırma performansına sahip olduğu belirlenmiştir. Görüldüğü üzere hem eğitim hem test verisi için 12 veri setinin yedisinde radyal tabanlı çekirdek fonksiyonu daha başarılı sınıflandırma performansına sahiptir. Eğitim ve test verileri için DVM'nin sınıflandırma performanslarının ortalaması dikkate alındığında wine ve WBC veri setleri için sigmoid çekirdek fonksiyonu, diğer 10 veri seti için radyal tabanlı çekirdek fonksiyonu en iyi sınıflandırma performansına sahip çekirdek fonksiyonları olduğu görülmüştür. Buradan elde edilen sonuçlardan sınıflandırma performansları incelendiğinde radyal tabanlı çekirdek fonksiyonun genel olarak daha başarılı sonuçlar verdiği kanısına varılabilir. Fakat sadece bu sonuçlardan bir fikre sahip olmak sakıncalı olabilir. Dolayısıyla algoritmanın sınıflandırma performansları açısından çekirdek fonksiyonları arasında farklılık olup olmadığı konusunda net bir yargıya sahip olabilmek için istatistiksel olarak test edilmesi (denen-

mesi) gerekmektedir (1).

2 DRY BEAN DATASET ÜZERİNDE ÇEKİRDEK FONKSİYONLARININ KARŞILAŞTIRILMASI

2.1 VERİ SETİ

Çalışmamda Kuru Fasulye veri setini kullandım. Bu veri setini (UCI Machine Learning Repository) çalışmasın da Selçuk Üniversitesindeki araştırmacılar oluşturmuş ve 2020 yılında açık erişime sunmuştur (2). 7 farklı tescilli kuru fasulyeye (Şeker, Barbunya, Bombay, Cali, Dermosan, Horoz, Sira) ait veri seti 13611 tane örnekten ve 17 tane öznitelikten oluşmaktadır. Bu öznitelikler aşağıdaki gibidir.

1) Area (Alan) : Fasulye bölgesinin alanı ve sınırları içindeki piksel sayısı.

2) Perimeter (Çevre) : Fasulye çevresi, kenarlığının uzunluğu olarak tanımlanır.

3) Major axis length (Ana eksen uzunluğu) : Bir çekirdekten çekilebilecek en uzun çizginin uçları arasındaki mesafe.

4) Minor axis length (küçük eksen uzunluğu) : Ana eksene dik dururken çekirdekten çekilebilen en uzun çizgi.

5) Aspect ration (görünüm) : Ana eksen uzunluğu ve küçük eksen uzunluğu arasındaki ilişkiyi tanımlar. Aşağıdaki şekilde hesaplanır.

$$Aspectration = \frac{(Majoraxislength)}{(Minoraxislength)}$$

6) Eccentricity (dışmerkezlik) : Bölge ile aynı momentlere sahip elipsin dışmerkezliği.

7) Convex area (dışbükey alan) : Bir fasulye tohumunun alanını içerebilen en küçük dışbükey çokgendeki piksel sayısı.

8) Equivalent diameter (eşdeğer çap) : Fasulye tohum alanı ile aynı alana sahip bir dairenin çapı. Aşağıdaki şekilde hesaplanır.

$$Equivalentdiameter = \sqrt{\frac{4 * (Area)}{\pi}}$$

9) Extent (Kapsam) : Sınırlayıcı kutudaki piksellerin fasulye alanına oranı.

10) Solidity (sağlamlık) : Dışbükeylik olarak da bilinir. Dışbükey kabuktaki piksellerin fasulyede bulunanlara oranı. Aşağıdaki şekilde hesaplanır.

$$Solidity = \frac{(Area)}{(Convexarea)}$$

11) Roundness (yuvarlaklık) : Aşağıdaki formülle hesaplanır.

$$Roundness = \frac{(4 * \pi * Area)}{(Perimeter)^2}$$

12) Compactness (Kompaktlık) : Bir nesnenin yuvarlaklığını ölçer. Aşağıdaki şekilde hesaplanır.

$$Compactness = \frac{(Equivalentdiameter)}{(Majoraxislength)}$$

13) ShapeFactor1 : Aşağıdaki formülle hesaplanır.

$$ShapeFactor1 = \frac{(Majoraxislength)}{(Area)}$$

14) ShapeFactor2 : Aşağıdaki formülle hesaplanır.

$$ShapeFactor2 = \frac{(Minoraxislength)}{(Area)}$$

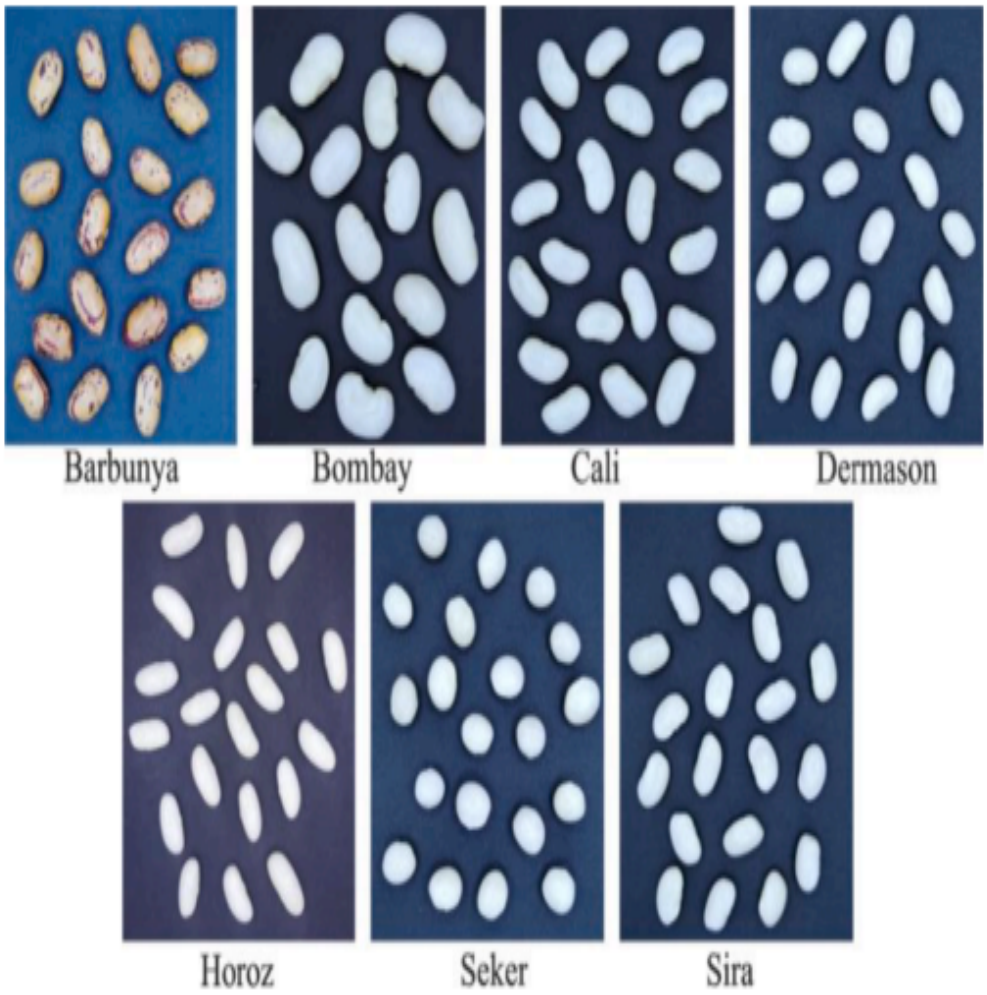
15) ShapeFactor3 : Aşağıdaki formülle hesaplanır.

$$ShapeFactor3 = \frac{(Area)}{\frac{(Majoraxislength)}{2} * \frac{(Majoraxislength)}{2} * \pi}$$

16) ShapeFactor4 : Aşağıdaki formülle hesaplanır.

$$ShapeFactor4 = \frac{(Area)}{\frac{(Majoraxislength)}{2} * \frac{(Minoraxislength)}{2} * \pi}$$

17) Class (Sınıf) : Seker, Barbunya, Bombay, Cali, Dermosan, Horoz ve Sira



Şekil 6: Kuru Fasulye Tohumlarının Örnek Resimleri

Kuru Fasulye veri seti eksik değerler içermemektedir. Veri tipi çok değişkenlidir. Özniteliklerin türleri; Kategorik, Gerçek ve Tamsayı şeklindedir.

Tablo 1: Kuru Fasulye Veri Setinin Dağılımı

NUMARA	KURU FASULYE TÜRÜ	*FASULYE ÇEŞİT SAYISI*	*FASULYE ÇEŞİT YÜZDELİĞİ (%)*
1	*ŞEKER*	*2027*	*14,89*
2	*DERMASON*	*2636*	*19,37*
3	*BOMBAY*	*522*	*3,84*
4	*BARBUNYA*	*1322*	*9,71*
5	*SIRA*	*1928*	*14,17*
6	*ÇALI*	*1630*	*11,98*
7	*HOROZ*	*3546*	*26,05*

Tablo 1 de kuru fasulye veri setinin dağılımı verilmektedir. Tablo 1 de kuru fasulye çeşitlerinin veri setinde ki adet sayısı ve yüzdelikleri verilmiştir. Bu tabloya göre en çok çeşit sayısı bulunan fasulye türü 3546 örnek ile Horoz fasulye olmuştur. En az çeşit sayısı bulunan fasulye türü ise 522 örnek ile Bombay fasulye olmuştur.

Veri setinde ki 16 özniteliğin aralarında ki ilişkisini görebilmek ve yorum yapabilmek için korelasyon matrisi Şekil 7 de ki gibi elde edilmiştir. Bunun için özelliklerimizin arasında ki korelasyon değeri $+1$ e ne kadar yakın ise iki değişkenimiz arasında pozitif korelasyon vardır , denir. Eğer iki değişkenimiz arasında ki korelasyon katsayısı -1 e ne kadar yakın ise iki değişkenimiz arasında negatif korelasyon vardır, denir.

Pozitif korelasyon değerine sahip iki değişkenin değerleri birlikte artar veya azalır. Yani iki değişken doğru orantılıdır.

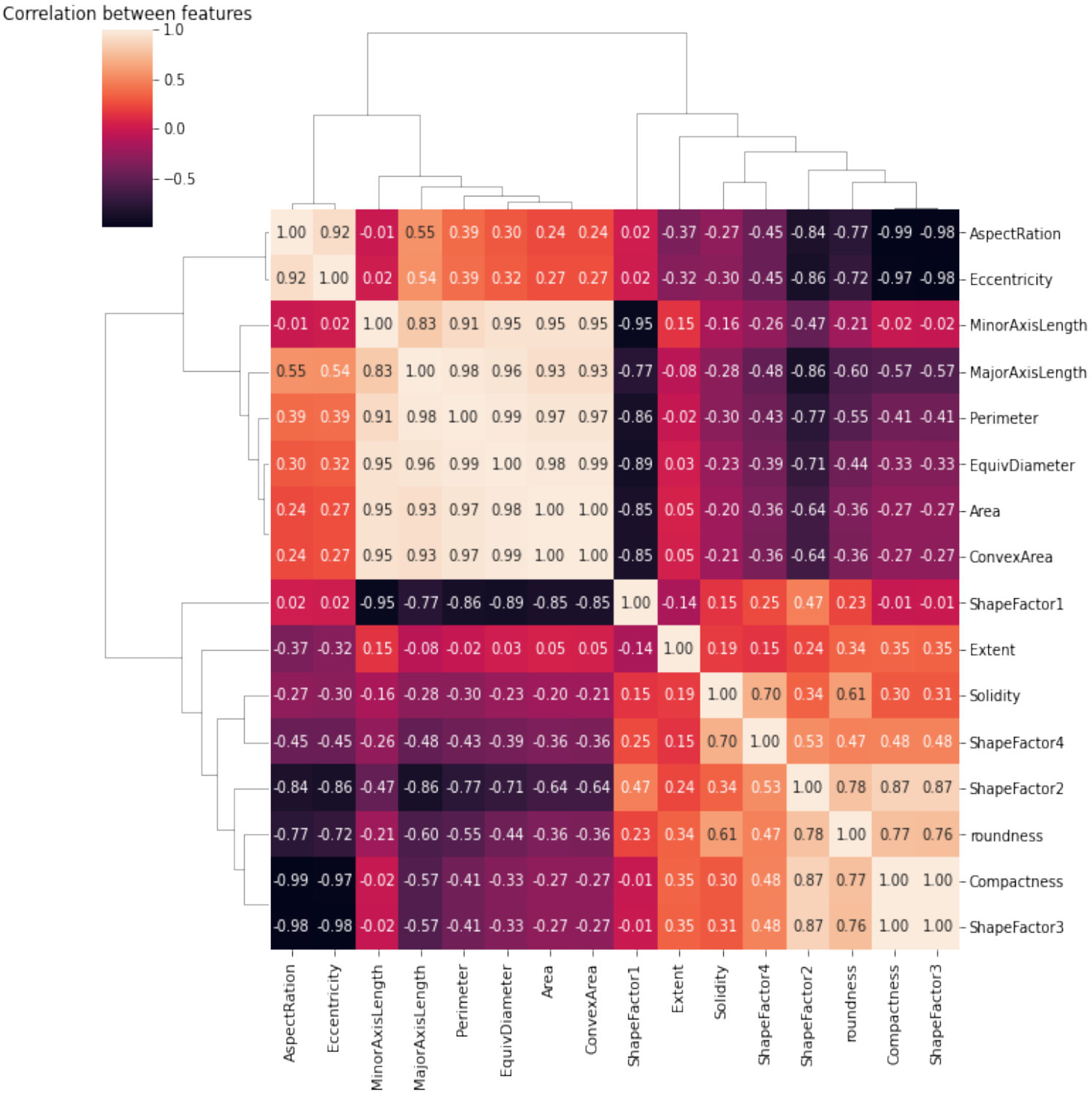
Negatif korelasyon değerine sahip iki değişkenin değerlerinden biri artarken, diğeri azalır. Yani iki değişken ters orantılıdır.

Korelasyon değerinin sıfıra yakın olması durumunda ise iki değişkenimiz arasında hiçbir ilişki olmadığını gösterir.

Şekil 7 den; Compactness ile ShapeFactor3 ve Area ile ConvexArea özniteliklerinin korelasyon değerinin 1 olduğu görülmektedir. Bu da bize bu özniteliklerin arsında pozitif korelasyon olduğunu gösterir ve bu değişkenlerin birlikte artar ya da azalır.

Compactness ile AspectRation öznitelikleri arasında ki korelasyon değeri -0.99 olduğu görülmektedir. Bu da bize bu özniteliklerin arasında negatif korelasyon olduğunu gösterir.

ShapeFactor3 ile AspectRation öznitelikleri arasında ki korelasyon değeri 0.02 olduğu görülmektedir. Bu da bize bu özniteliklerin arasında bir ilişki olmadığını gösterir.



Şekil 7: Korelasyon Matrisi

2.2 VERİ DENGESİZLİĞİNİN GİDERİLMESİ

Sınıflandırma problemlerinde kullanılan veri setlerinde her sınıfta hemen hemen aynı oranda örneklerin olması beklenir. Ancak bu durumun olmadığı zamanlarda veri setinde bir dengesizlik meydana gelir ve bu yüzden sınıflandırma yapılırken örneğin daha fazla olduğu tarafa doğru bir yönelim gerçekleşir. Bununla birlikte azınlık olan sınıfa, doğru ve başarılı bir sınıflandırma yapılamaz. Bu durum istenen bir sonuç değildir. Sınıflandırıcının her bir sınıf için yüksek başarı elde etmesi amaçlanmaktadır (3).

Veri setinde oluşan bu dengesizliği giderebilmek için kullanılabilecek çeşitli yöntemler bulunmaktadır. SMOTE (Synthetic Minority Over Sampling Technique) metodu bu sorunu gidermek için uygulanabilecek yöntemlerden birisidir.

Chawla et al., (2002), yaptıkları bir çalışmada veri setinde bulunan dengesizlik problemine çözüm bulabilmek amacı ile SMOTE algoritmasını önermişlerdir. Bu algoritma, azınlık olarak bulunan sınıfa ait kayıtların sayısını artırmak için sentetik veri üreterek veri setinde bulunan dengesizliği ortadan kaldırmaya yönelik bir yöntemdir.

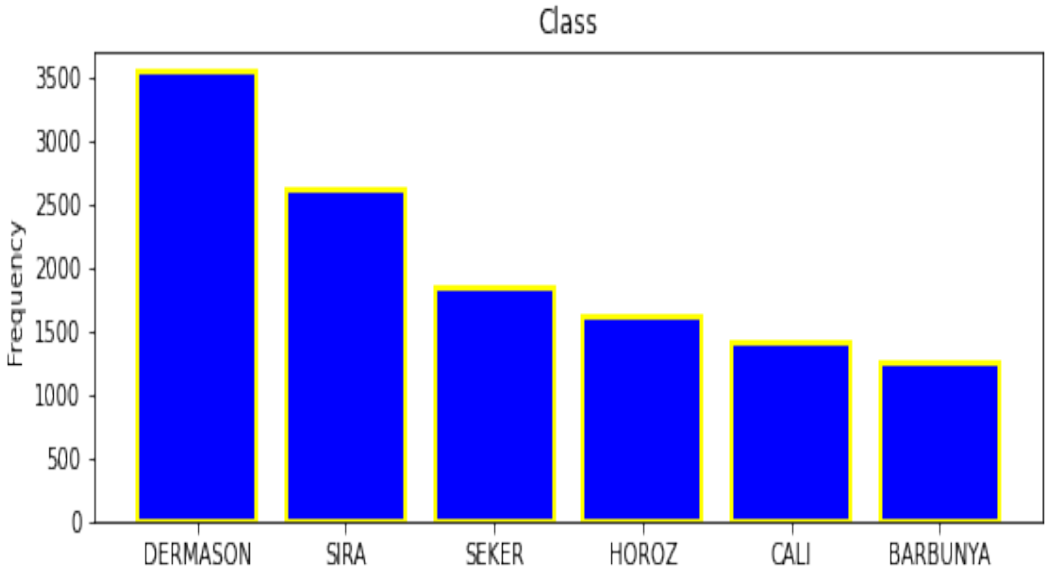
Sentetik veri üretimi aşağıdaki şekilde gerçekleşmektedir (4):

1) İncelenen özellik vektörü ile en yakın komşusu arasındaki fark hesaplanır.

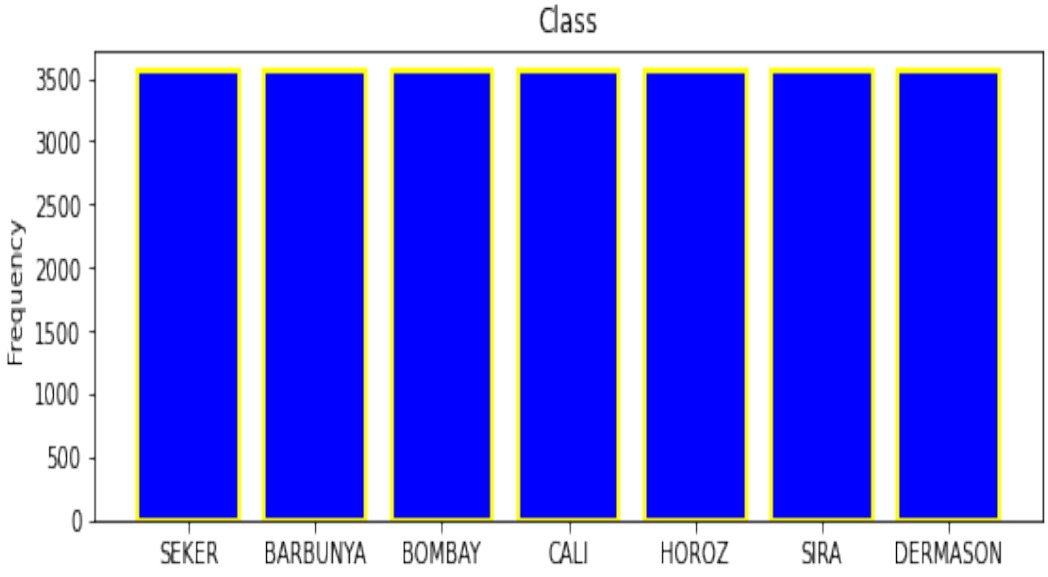
2) Bu fark 0 ile 1 arasında rastgele bir sayı ile çarpılır ve söz konusu olan özellik vektörüne eklenir.

3) Bu durum, iki belirli özellik arasındaki çizgi parçası boyunca rastgele bir noktanın seçilmesine sebep olur.

4) Bu yaklaşım, azınlık sınıfının karar bölgesini daha genel olmaya zorlamaktadır, yani yapay örneklemler oluşturmaktadır.



Şekil 8: SMOTE Öncesi Veri Dağılımı



Şekil 9: SMOTE Sonrası Veri Dağılımı

Şekil 8 den görüldüğü üzere kuru fasulye veri setinin Class sınıfına ait verileri dengesiz bir şekilde dağılmaktadır. Bu dengesizliği, veri seti üzerinde yapılacak olan çalışmaların sonuçlarını yanlış yönlendirilecek etkilere sebep olabileceğinden bu problemi çözmek için SMOTE algoritması kullanılmıştır. Python' un Imbalanced-Learn kütüphanesinde olan SMOTE fonksiyonu ile örneklem

arttırma yapılmıştır. Böylece SMOTE algoritması ile Şekil 9 dan görülmek üzere Class sınıfına ait verilerin hepsi eşitlenmiştir.

2.3 ÖZELLİK ÖLÇEKLENDİRME

Sınıflandırma problemlerinde genellikle aynı veri kümesinde farklı türde değişkenler vardır. Bu problemin üstesinden gelmek için, veri ön işleme adımında, bağımsız değişkenlere veya verilerin özelliklerine yeniden ölçeklendirme özelliği tekniğini uygulamamız gerekir.
Özellik ölçeklendirmeyi uygulamanın amacı, özelliklerin hemen hemen aynı ölçekte olduğundan emin olmak ve böylece her bir özelliğin eşit derecede önemli düzeyde olmasını sağlamaktır.
Standardizasyon, değişken (özellik) sütunlarının ortalama değeri 0 ve standart sapması 1 olacak şekilde standart normal dağılım oluşturmaktır.
Genel olarak, ölçeklerdeki bazı farklılıkları korurken verileri normalleştirmek istiyorsanız, (çünkü birimler farklı kalır) minimum-maks normalleştirmeyi kullanmak ve ölçekleri karşılaştırılabilir hale getirmek istiyorsanız (standart sapmalar aracılığıyla) standardizasyonu kullanmak bu uygulamanın temel kuralıdır.

2.4 VERİ SETİNDEN ÖNEMLİ ÖZELLİK SEÇİMİ

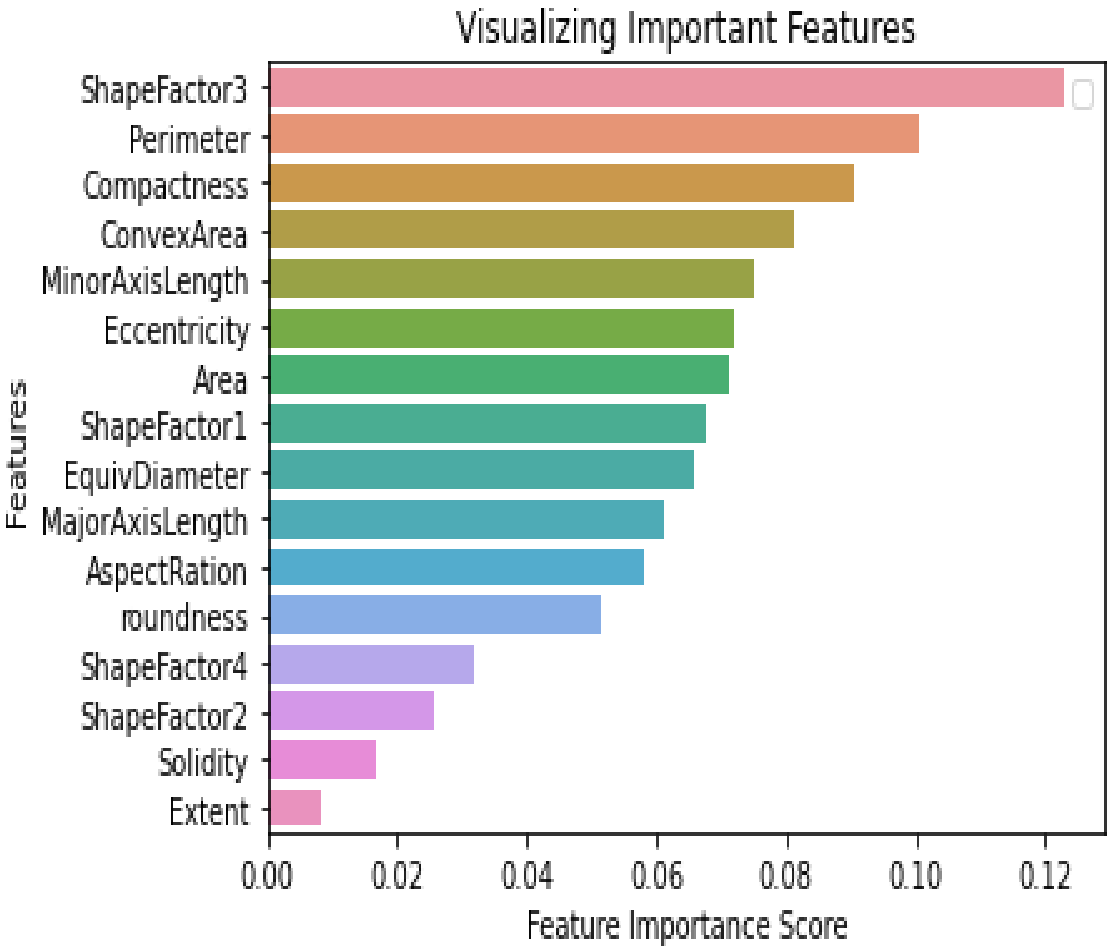
Veri setinden önemli özellik seçme işlemi birçok yol ile yapılır. Veri setinden önemli özellikleri seçme işlemini Random Forest algoritmasını kullanarak gerçekleştirdim. Çünkü Random Forest modelinin diğer bir özelliği bize özniteliklerin ne kadar önemli olduğunu vermesi. Bir özniteliğin önemli olması demek o özniteliğin bağımlı değişkendeki varyansın açıklanmasına ne kadar katkı yaptığıyla alakalı. Random forest algoritmasına x sayıda öznitelik verip en faydalı y tanesini seçmesini isteyebiliriz ve istersek bu bilgiyi istediğimiz başka bir modelde de kullanabiliriz. Random Forest algoritması bu işlemi her bir özelliğin tahmin üzerindeki nispi önemini ölçerek gerçekleştirir.
Random Forest sınıflandırma kuralları oluşturulurken doğrudan değişken seçimini gerçekleştirir. Değişken önemliliğinin bulunmasındaki en önemli amaçlar; model performansını geliştirerek aşırı uyumu yok etmeye çalışmak ve veri setini türeten sürecin altında yatan süreçleri daha iyi anlamaya çalışmaktır (5).
Değişken önemliliği birbirine paralel sonuçlar veren iki yöntem ile hesaplanabilir. Bunlar; Gini önemliliği ve permütasyona dayalı değişken önemliliğidir.
Gini önemliliği, doğrudan Random Forest ağaçları oluşturulurken kullanılan Gini indeksinden elde

edilir. Gini indeksi bir düğüme atanmış örneklemin karışıklık ya da eşitsizlik seviyesini ölçer. Örneğin, iki sınıflı bir sınıflandırma probleminde p; k düğümünde yer alan pozitif gözlemlerin oranını ve 1-p de negatif gözlemlerin oranını göstersin. Bu durumda k düğümünde yer alan Gini indeksi aşağıdaki gibi hesaplanır:

$$G_k = 2p(1-p)$$

Bir düğüm ne kadar saflaştırılırsa, Gini değeri de o kadar küçülür.
Random Forest yönteminde v değişkeninin önem derecesi aşağıdaki sıralama ile bulunur. Öncelikle OOB (torba dışı gözlemler) gözlemleri ağaçtan aşağı bırakılır ve tahmin edilen değerler belirlenir. Daha sonra ise OOB' de yer alan diğer tahminci değişkenler sabit olmak koşulu ile v değişkenine ait gözlem değerleri rasgele karıştırılır. Elde edilen yeni OOB veri seti ağaçtan aşağı bırakılır ve tahmin edilen değerler belirlenir. Bu işlem sonucunda her gözlem için iki tane tahmin değeri elde edilmiş olur. Orijinal OOB ile elde edilen doğru tahmin sayısından, değiştirilmiş OOB ile elde edilen doğru tahmin sayısı çıkartılarak bir fark elde edilir. Bu işlem tüm ormana uygulanarak ormandaki ağaç sayısı kadar fark elde edilir ve bu farkların ortalaması hesaplanır. Tüm ağaçların birbirinden bağımsız olduğu ve elde edilen fark değerlerinin normal dağıldığı varsayımı altında v değişkeni için z skor değeri hesaplanır. Bu skor değeri; farklar ortalamasının farkların standart hatasına oranlanması ile hesaplanır. Ağaçta yer alan her v değişkeni için skor değerler elde edilir. Elde edilen skor değerlerine göre değişkenlerin önemlilik dereceleri kıyaslanarak bir sıralama belirlenmiş olur(6, 7, 8).

OOB (torba dışı gözlemler) puanı, bir makine öğrenimi modeli olan Random Forest gibi topluluk modelleri için bir performans ölçümüdür. Torba dışı gözlemler olarak adlandırılan modelin eğitiminde kullanılmayan gözlemler kullanılarak hesap edilir. Bu örnekler, modelin OOB puanı olarak bilinen performansının tarafsız bir tahminini sağlamak için kullanılır.



Şekil 10: Veri Setinin Önemli Özellikleri

Random Forest kullanarak veri setinden seçilen önemli özellikler Şekil 10 dan aşağıdaki gibi bulunmaktadır.
ShapeFactor3, Perimeter, Compactness, ConvexArea, MinorAxisLength, Eccentricity.

Başka bir ifadeyle Random Forest algoritması, birden fazla karar ağacı üreterek sınıflandırma işlemi esnasında sınıflandırma değerini yükseltmeyi hedefler. Random Forest algoritması birbirinden bağımsız olarak çalışan birçok karar ağacının bir araya gelerek aralarından en yüksek puan alan değerin seçilmesi işlemidir. Ağaç sayısı arttıkça kesin bir sonuç elde etme oranımız artmaktadır.

Random Forest algoritması, elinde yeterli miktarda ağaç varsa aşırı öğrenme sorununu azaltır. Az oranda bir veri hazırlığına ihtiyaç duyar.

Günlük hayattan bir örnek ile daha iyi anlaşılmasını sağlarsak, hastalandığımız zaman doktora gideriz. Doktora, hastalığımızı tespit etmesi için hastalığın sebep olduğu belirtileri söyleriz. Doktor belirtilere göre çeşitli tahliller yapar ve başka doktorlardan hastalığı teşhis etmek için görüş ister. Doktorların görüşleri ve yapılan tahlillerin sonucuna göre hastalık teşhis edilir ve hasta olan kişi önerilen tedavi doğrultusunda hastalığı atlatır. Burada her bir doktor bir karar ağacını temsil etmektedir. Hasta olan kişi, doktorlardan oluşan çoğunluğun verdiği karar doğrultusunda hangi tür hastalığa sahip olduğunu öğrenir. Bu sonuca göre tedavisi gerçekleştirilir. Random Forest algoritması da bu mantıkla çalışan bir yapıya sahiptir.

2.5 VERİ SETİNİN TRAİN VE TEST AYRIMI

Makine öğrenmesi modelinin başarısını test etmek (ölçmek) amacıyla veri setini eğitim ve test olarak iki parçaya böleriz.
Train Set (Eğitim seti) : Modelin eğitildiği veri kümesidir.
Test Set (Test seti) : Eğitilen modelimizi test etmek için kullanılan veri kümesidir.
Eğitim setimiz büyüdükçe modelimiz daha iyi öğrenecektir.
Modelimiz de test size 0.2 olarak alınmıştır.

2.6 DESTEK VEKTÖR MAKİNESİ(SVM)

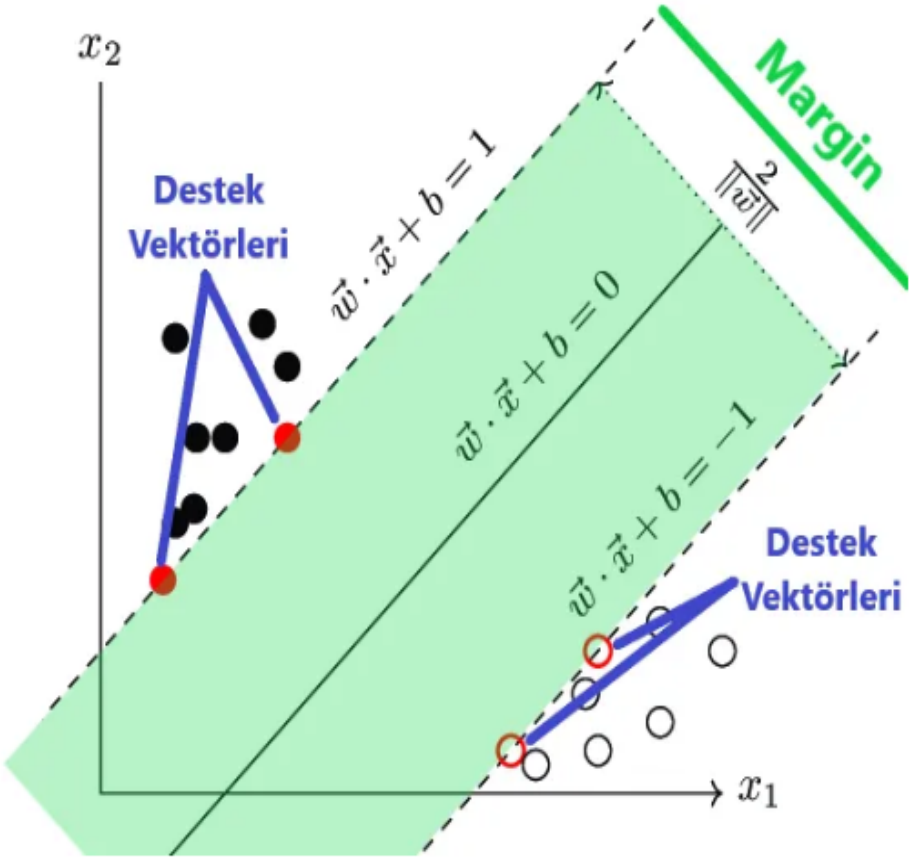
Destek Vektör Makinesi, Vapnik - Chervonenkis tarafından geliştirilen, sınıflandırma problemleri başta olmak üzere kümeleme ve regresyon problemlerinde kullanılan bir gözetimli öğrenme algoritmasıdır. Destek Vektör Makineleri, temel olarak iki sınıfa ait verileri birbirinden en uygun şekilde ayırmak için kullanılan bir algoritmadır. Bunun için karar sınırları ya da diğer bir ifadeyle hiper düzlemler belirlenir.

Destek Vektör Makinelerinin avantajları arasında;

- Yüksek boyutlu uzaylarda etkili olması,
- Boyut sayısının, örneklem sayısından fazla olduğu durumlarda etkili olması,

- Karar fonksiyonunda bir takım eğitim noktaları kullanılması ("support vectors"). Dolayısıyla da belleğin verimli bir şekilde kullanılmış olması,
- Çok yönlü olması. Yani, karar fonksiyonu için çok farklı çekirdek fonksiyonları ("kernel functions") kullanılabilmesi, gibi durumlar mevcuttur.

Destek Vektör Makineleri, veri setinin doğrusal olarak ayrılabilme ve doğrusal olarak ayrılamama durumlarına göre ikiye ayrılmaktadır.



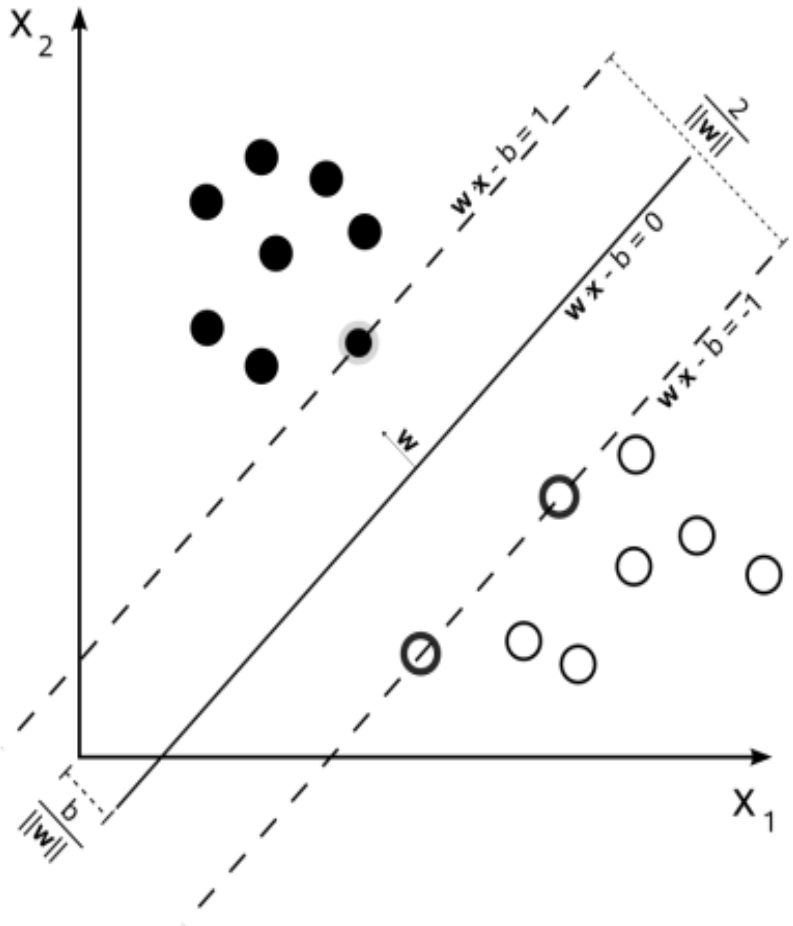
Şekil 11: Destek Vektörleri

Şekil 11 de siyahlar sınıfı ve beyazlar sınıfı olmak üzere iki farklı sınıf mevcuttur. Sınıflandırma problemlerindeki amaç gelecek yeni bir verinin hangi sınıfta yer alacağını karar vermektir. Bu sınıflandırmayı yapabilmek için iki sınıfı ayıran bir doğru çizilir ve bu doğrunun 1'i arasında kalan yeşil

bölgesine Margin denir. Margin ne kadar geniş ise iki veya daha fazla sınıf o kadar iyi ayrıştırılır.

2.6.1 DOĞRUSAL DESTEK VEKTÖR MAKİNESİ

Doğrusal destek vektör makinelerinde veri grupları doğrular (çizgiler) ile kolay bir şekilde ayrılıp, veriler düzlem-hiper düzlem ile sınıflandırılmaktadır.
Veri setini ikiye ayıran doğru karar doğrusu olarak isimlendirilmektedir. Sonsuz tane karar doğrusu çizebilme imkanı mevcut olsada önemli olan optimal yani en uygun karar doğrusunu belirlemektir.

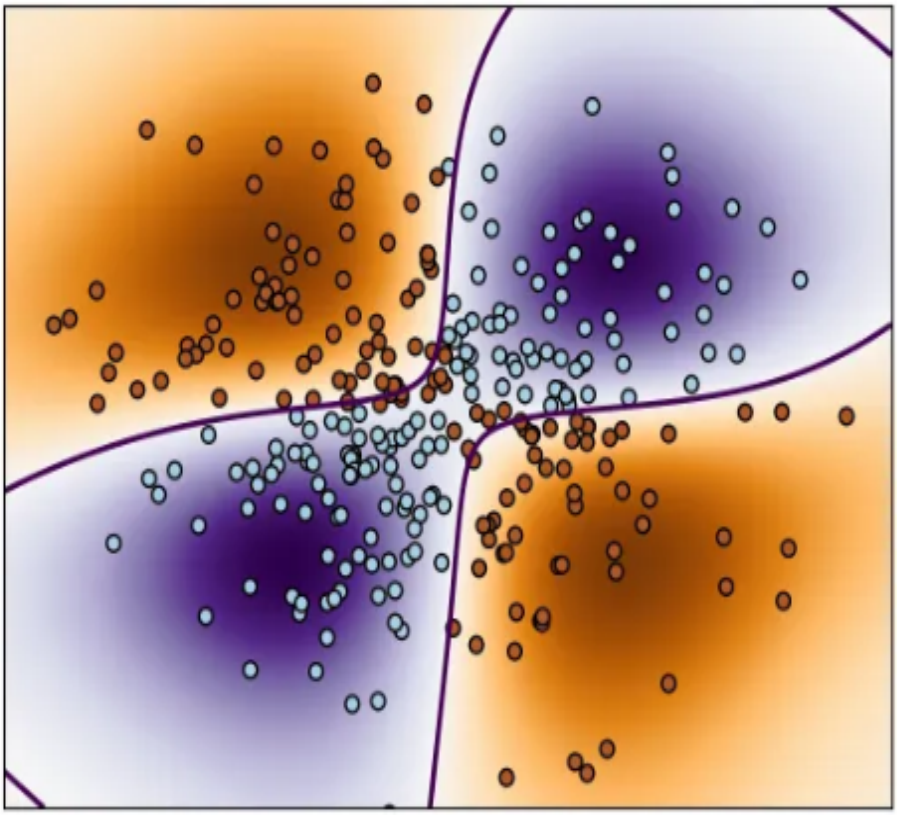


Şekil 12: Doğrusal Destek Vektör Makinesi

2.6.2 DOĞRUSAL OLMAYAN DESTEK VEKTÖR MAKİNESİ

Doğrusal olmayan bir veri kümesinde SVM'ler doğrusal bir hiper-düzlem çizemez. Bu nedenle çekirdek hilesi olarak adlandırılan kernel trick kullanılır. Düşük boyutlar karmaşık veri setlerini açıklamada yeterli olmayabilir. Boyutu arttırsak işlemler artacağı için çok uzun sürer. İşte kernel trick burada devreye giriyor. Elimizdeki koordinatları belirli çekirdek fonksiyonları ile çarparak daha çok anlamlı hale getirebiliriz.

Çekirdek yöntemi, doğrusal olmayan verilerde makine öğrenimini yüksek oranda arttırmaktadır.



Şekil 13: Doğrusal Olmayan Destek Vektör Makinesi

2.7 ÇEKİRDEK FONKSİYONLARI

SVM algoritmaları, çekirdek olarak tanımlanan bir dizi matematiksel işlev kullanır. Çekirdeğin işlevi, verileri girdi olarak almak ve gerekli forma dönüştürmektir. Farklı SVM algoritmaları, farklı türde çekirdek işlevleri kullanır. Çekirdek fonksiyonları sınıflandırma ve regresyon problemlerinde bir benzerlik fonksiyonu olarak görev yapmaktadır. Bu fonksiyonlar makine öğrenmesi algoritmalarına çeşitli esnek özellikler kazandırır. Böylelikle sınıflandırma ve regresyon problemlerinde bu esnek özellikleri nedeni ile daha yüksek performans sonuçları elde edilebilmektedir.

Bu çalışmada doğrusal (linear), Gaussian radyal tabanlı fonksiyon (rbf), polinomiyal ve sigmoid çekirdek fonksiyonları kullanılmıştır.En çok kullanılan çekirdek işlevi türü RBF'dir . Çünkü tüm x ekseni boyunca lokalize ve sonlu tepkiye sahiptir.

- Doğrusal (Linear) Çekirdek Fonksiyonu:

Doğrusal çekirdek fonksiyonu, en basit çekirdek fonksiyonudur. İç çarpımı ile ifade edilir ve doğrusal çekirdek fonksiyonunun denklemi aşağıda verilmektedir.

$$K(x_i, x_j) = x_i^T x_j$$

- Polinom (Poly) Çekirdek Fonksiyonu:

Polinom çekirdek fonksiyonu, durağan olmayan bir çekirdek fonksiyonudur. Polinom çekirdeği fonksiyonu, genellikle tüm eğitim verilerinin normalleştirildiği problemler için kullanılır ve denklemi aşağıda verilmektedir.

$$K(x_i, x_j) = (1 + x_i^T x_j)^d$$

- Radyal Tabanlı Gauss Çekirdek Fonksiyonu (RBF):

Gauss çekirdek fonksiyonu radyal taban fonksiyonu çekirdeğine bir örnektir. RBF çekirdek fonksiyonun denklemi aşağıda verilmektedir.

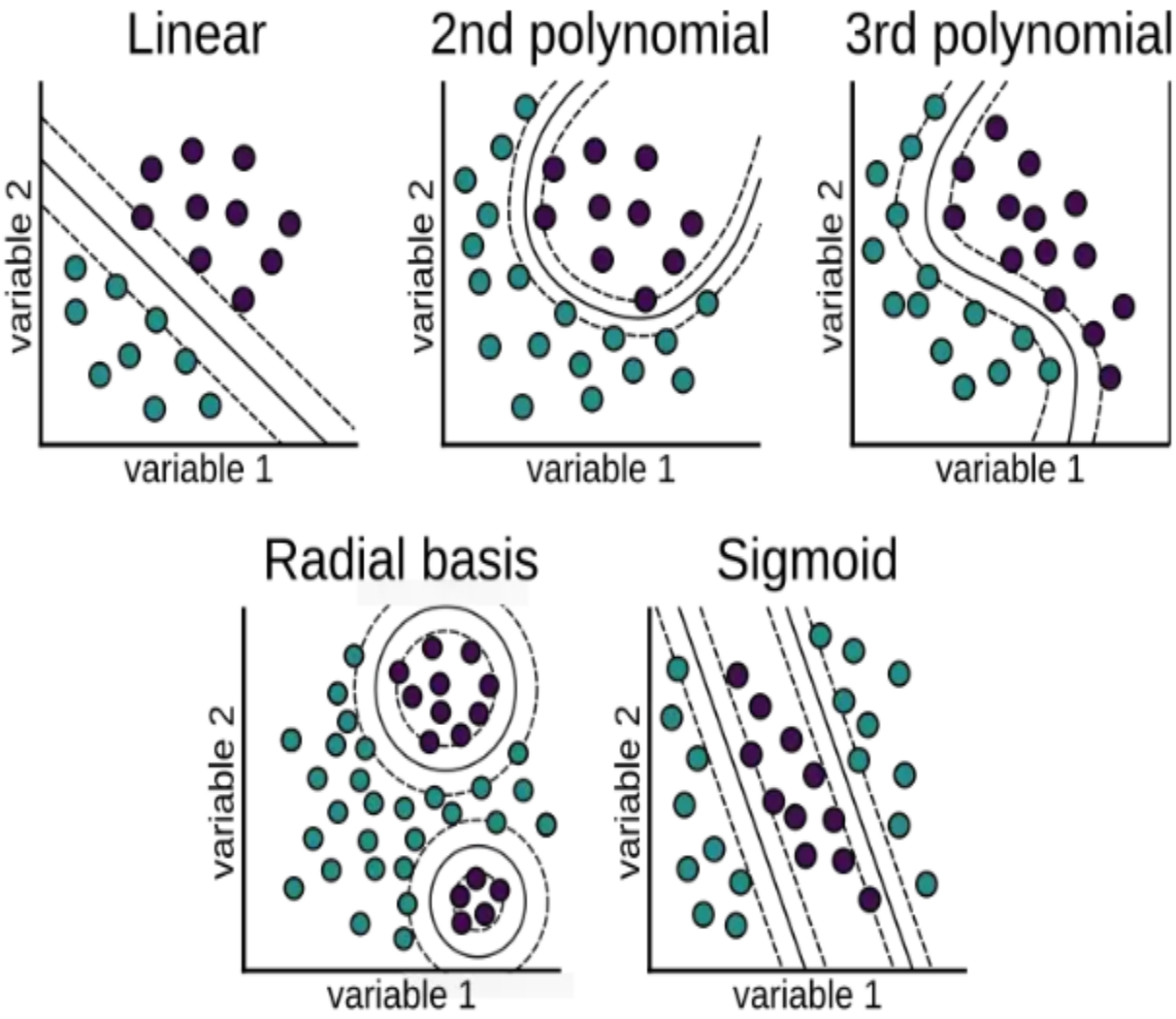
$$k(x, y) = \exp(-\gamma \left\| x - y \right\|^2)$$

Ayarlanabilir parametre olan gamma değeri, çekirdeğin performansında önemli bir rol oynar ve eldeki soruna göre dikkatle seçilmelidir. Aşırı tahmin söz konusu olduğunda, üstel ifade yaklaşık olarak doğrusallaşır ve yüksek boyutlu projeksiyon doğrusal olmayan yapısını kaybetmeye başlar.

- Sigmoid Çekirdek Fonksiyonu:

Sigmoid çekirdek fonksiyonu, birçok uygulamada iyi performans göstermektedir. Sigmoid çekirdeğinde α eğim ve c kesme sabiti olmak üzere iki ayarlanabilir parametresi vardır. α için ortak bir değer (1 / N) kullanılabilir ve N değeri veri boyutunu ifade etmektedir. Sigmoid çekirdek fonksiyonunun denklemi aşağıda verilmektedir.

$$K(x_i, x_j) = \tanh(\alpha x_i^T x_j + c)$$



Şekil 14: Çekirdek Fonksiyonlarının İşlevleri

Tablo 2: Çekirdek Fonksiyonlarının Karşılaştırılması

ÇEKİRDEK FONKSİYONLARI	SMOTE TEKNİĞİ KULLANILMADAN ELDE EDİLEN SONUÇLAR	SMOTE TEKNİĞİ KULLANILARAK ELDE EDİLEN SONUÇLAR
DOĞRUSAL	0,926829268292683	0,9302240176276166
RBF	0,93270643549809	0,9383033419023136
POLYY	0,9338818689391714	0,937201615864855
SİGMOİD	0,9191889509256538	0,9162688211531399

Tablo 2 de çekirdek fonksiyonlarının karşılaştırılması verilmiştir.

Tablo 2 deki sonuçlar ışığında şu çıkarımları yapabiliriz:

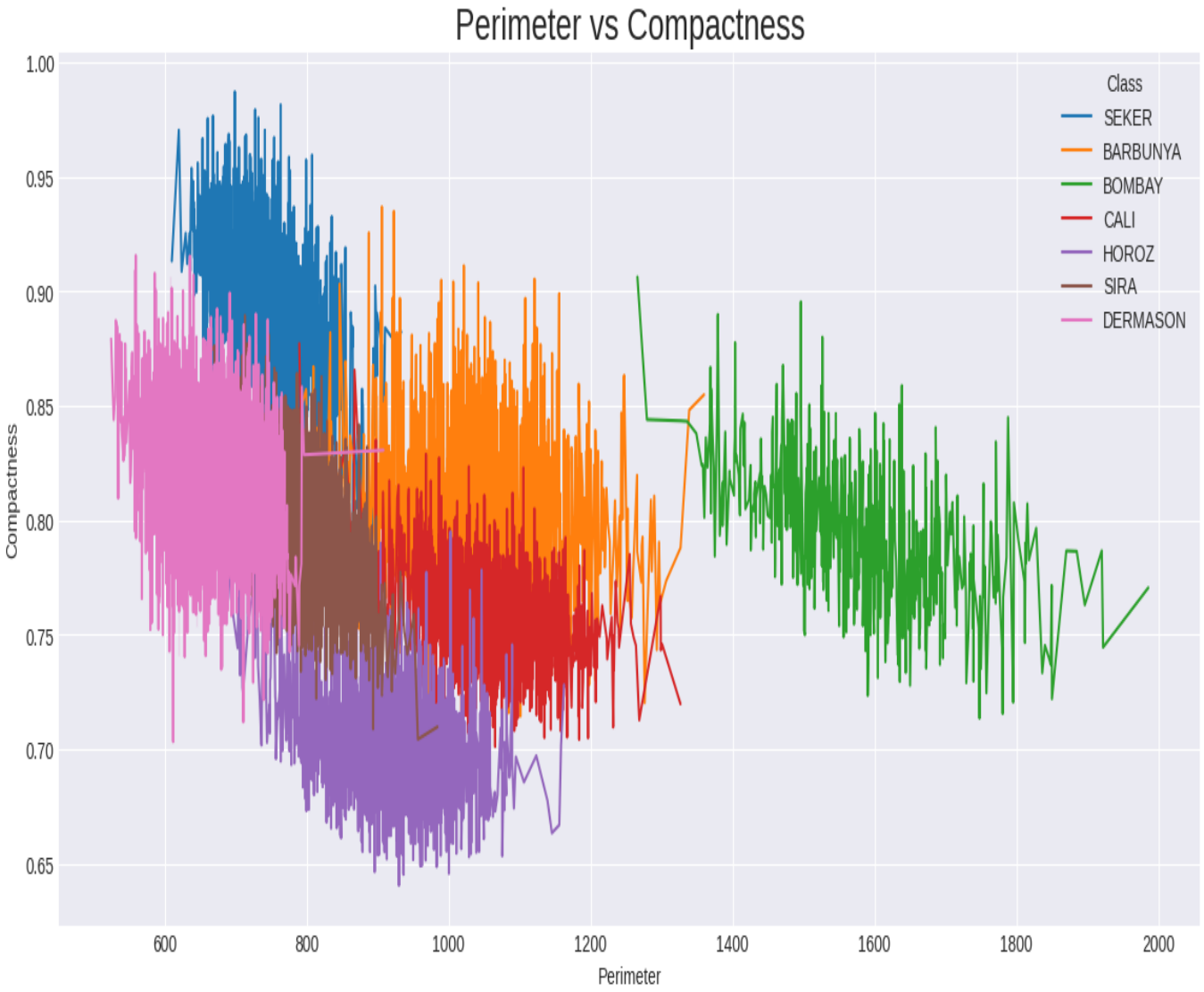
- Doğrusal çekirdek fonksiyonu ile kurulan modelimiz, veri noktalarımızın % 93,02 sini doğru bir şekilde sınıflandırabilmiştir.
- RBF çekirdek fonksiyonu ile kurulan modelimiz, veri noktalarımızın % 93,83 ünü doğru bir şekilde sınıflandırabilmiştir. Ayrıca RBF çekirdek fonksiyonu en iyi sonucu $\gamma = 0.09$ ve $C = 10$ değerleri için vermektedir.
- Poly çekirdek fonksiyonu ile kurulan modelimiz, veri noktalarımızın % 93,72 sini doğru bir şekilde sınıflandırabilmiştir. Ayrıca poly çekirdek fonksiyonu en iyi sonucu $\gamma = 1$ ve degree=2 değerleri için vermektedir.
- Sigmoid çekirdek fonksiyonu ile kurulan modelimiz, veri noktalarımızın % 91,62 sini doğru bir şekilde sınıflandırabilmiştir. Ayrıca sigmoid çekirdek fonksiyonu en iyi sonucu $\gamma = 0.01$ değeri için vermektedir.

Bu sonuçlara bakarak en yüksek doğruluğu elde ettiğimiz çekirdek fonksiyonu % 93,83 ile RBF çekirdek fonksiyonu olmuştur. RBF çekirdek fonksiyonundan sonra en yüksek doğruluğu elde ettiğimiz çekirdek fonksiyonu Poly çekirdek fonksiyonu olmuştur. Burada en iyi sonuçları SMOTE tekniği kullanarak elde ettiğimiz de görülmektedir.

Bu tablodan SMOTE tekniğinin model çalışması üzerindeki olumlu etkisini de elde ettiğimiz sonuçlardan görmekteyiz.

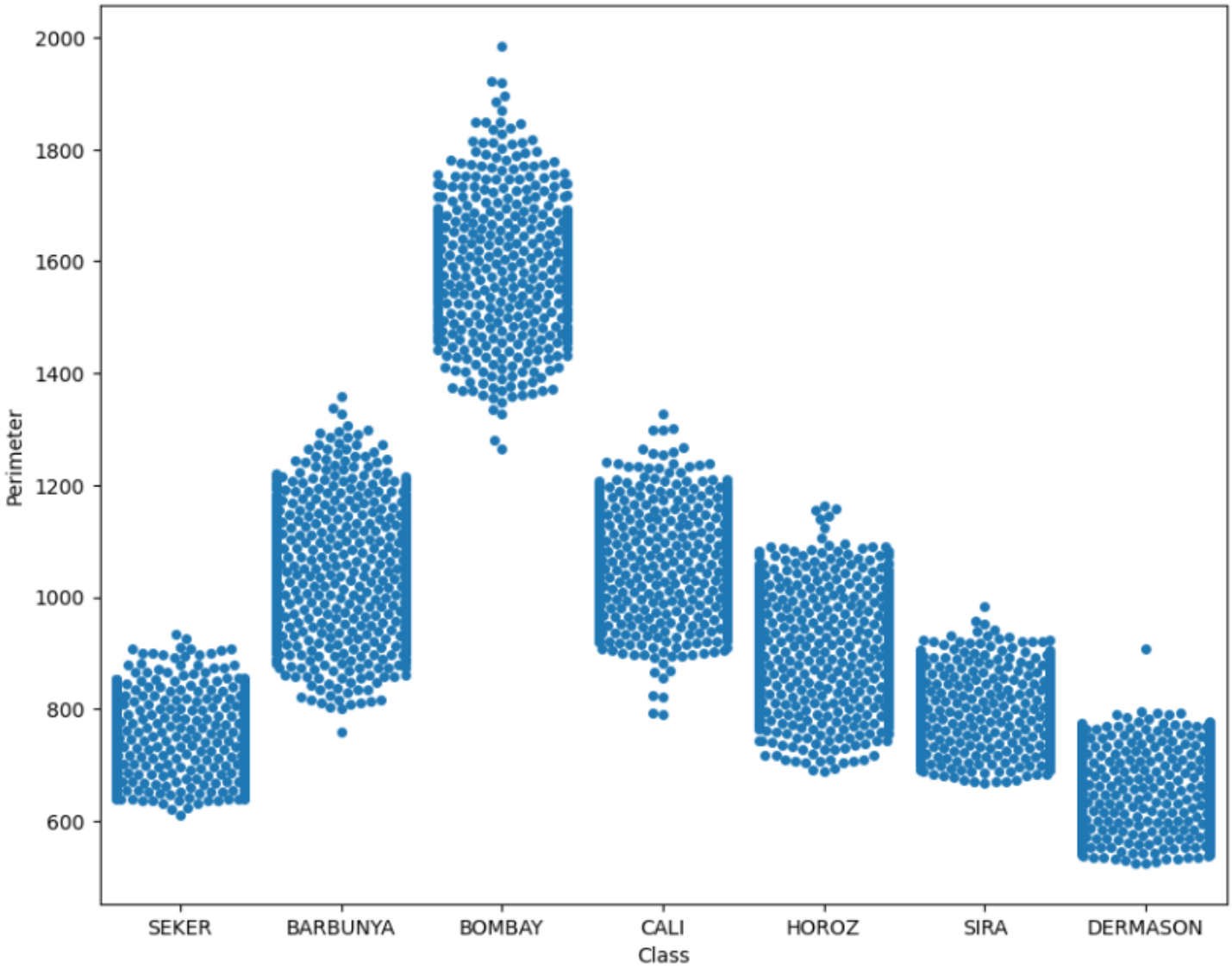
3 VERİ GÖRSELLEŞTİRME

Şekil 15: Class Değişkenine Ait Verilerin Perimeter ve Compactness Değişkenine Göre Dağılım Grafiği



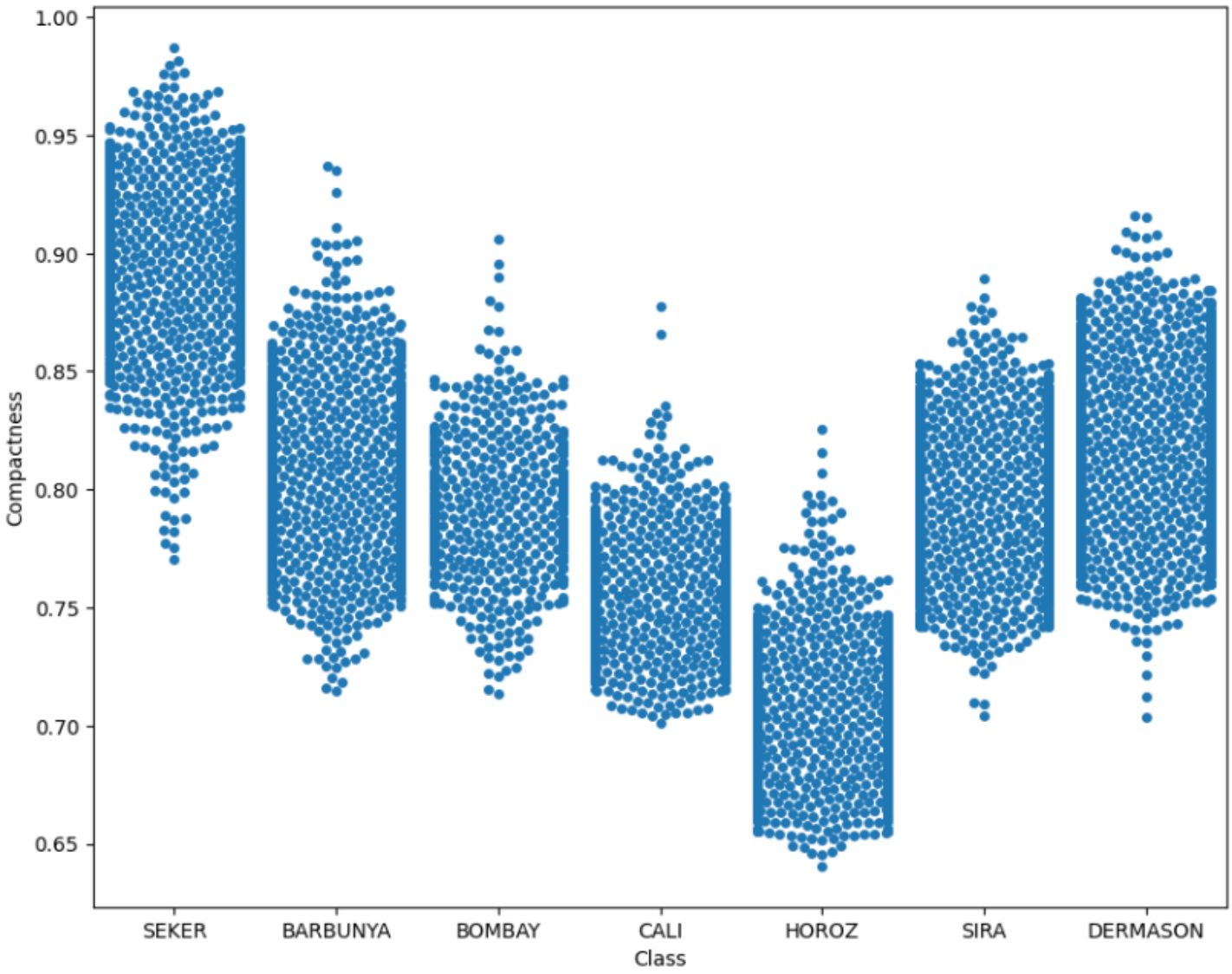
Şekil 16: Class Değişkenine Ait Verilerin Perimeter ve Compactness Değişkenine Göre Çizgi Grafiği

Şekil 15 de Class değişkenine ait fasulye türleri verilerinin Perimeter ve Compactness değişkenine göre dağılım grafiği verilmektedir. Şekil 16 da Class değişkenine ait fasulye türleri verilerinin Perimeter ve Compactness değişkenine göre çizgi grafiği verilmektedir. Kuru fasulyelerden Compactness değeri en yüksek olan Şeker fasulye iken en az olan Horoz fasulyedir. Perimeter değeri en yüksek olan Bombay fasulye iken en az olan Dermason fasulyedir.



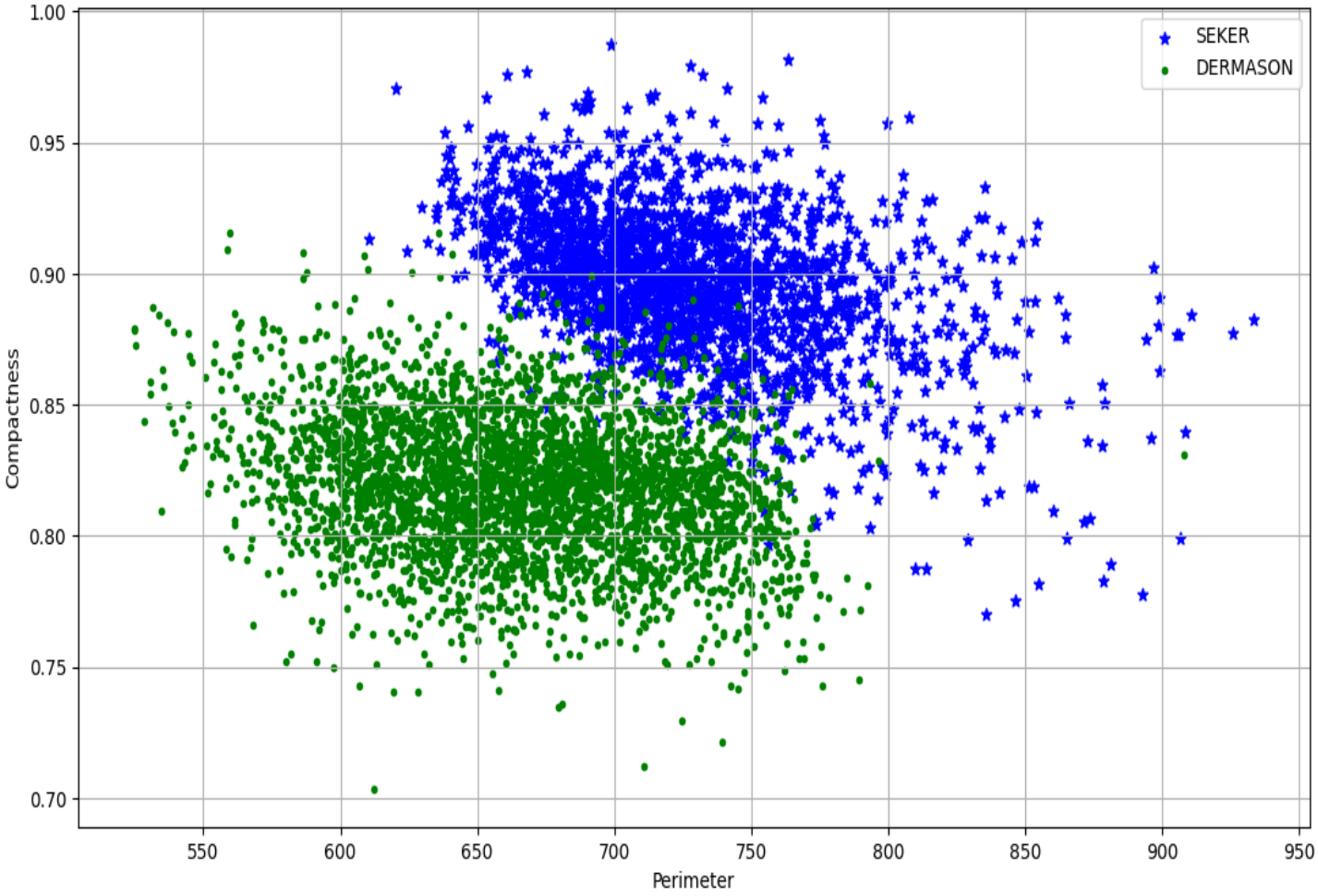
Şekil 17: Class Değişkenine Ait Verilerin Perimeter Değişkenine Göre Dağılımı

Şekil 17 de Class değişkenine ait fasulye türleri verilerinin Perimeter değişkenine göre dağılımı verilmektedir.
Bombay fasulyenin Perimeter değeri diğer fasulye türlerine göre en yüksek olan fasulye türüdür.
Dermason fasulyenin Perimeter değeri diğer fasulye türlerine göre en düşük olan fasulye türüdür.
Perimeter, fasulye çevresi olarak tanımlandığından dolayı en yüksek değeri Bombay fasulyenin ve en az değeri Dermason fasulyenin alması doğru bir gözlemdir.
Bu tür grafikler noktaların dağılımını da vermektedir ve aynı zamanda kategorik değişkenler üzerinde sayısal dağılımları grafik haline getirmektedir.



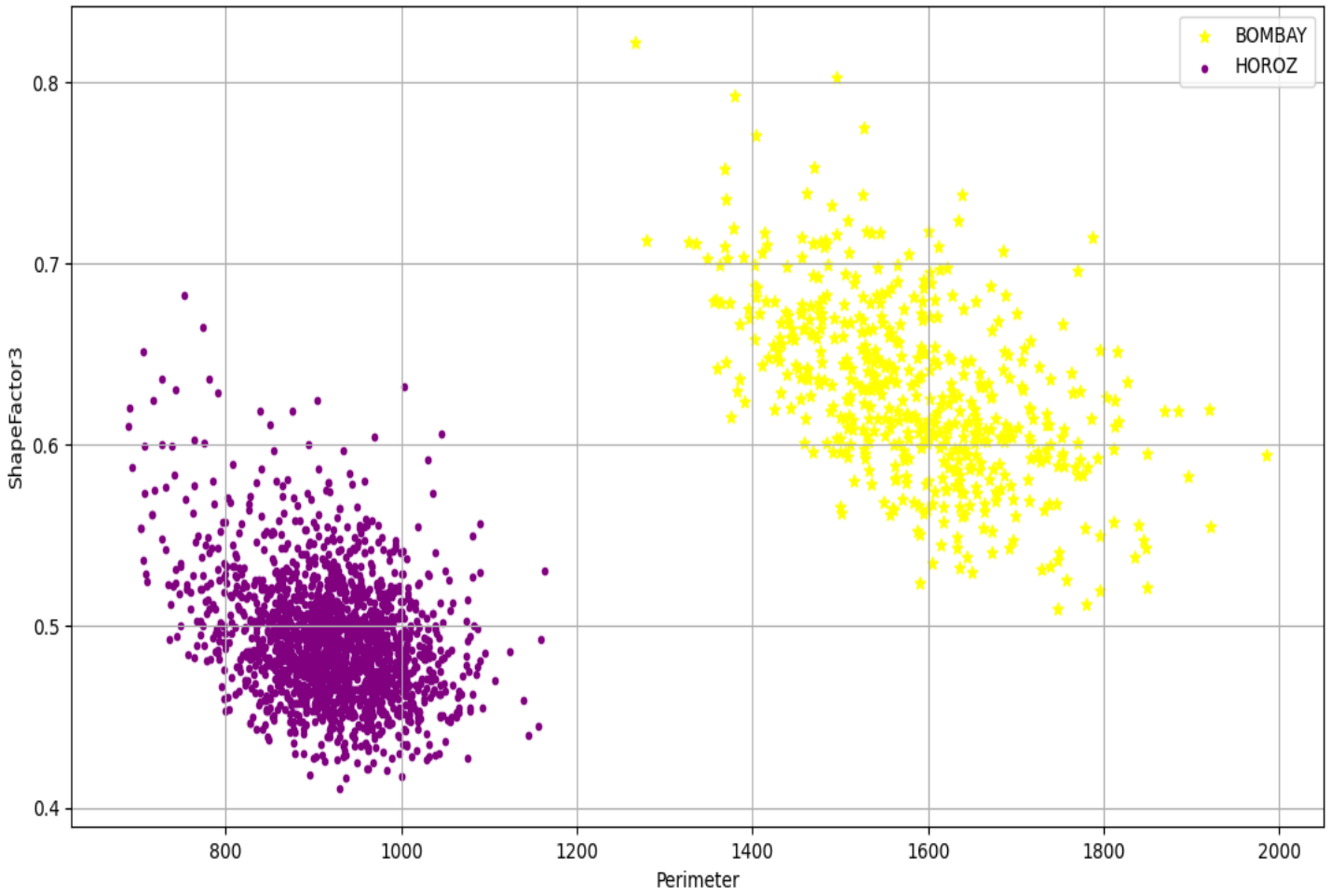
Şekil 18: Class Değişkenine Ait Verilerin Compactness Değişkenine Göre Dağılımı

Şekil 18 de Class değişkenine ait verilerin Compactness değişkenine göre dağılımı verilmektedir. Compactness, bir nesnenin yuvarlaklığının ölçümü olduğundan en yüksek değeri Şeker fasulyenin en az değeri Horoz fasulyenin alması doğru bir gözlemdir.
Bu tür grafikler noktaların dağılımını da vermektedir ve aynı zamanda kategorik değişkenler üzerinde sayısal dağılımları grafik haline getirmektedir.



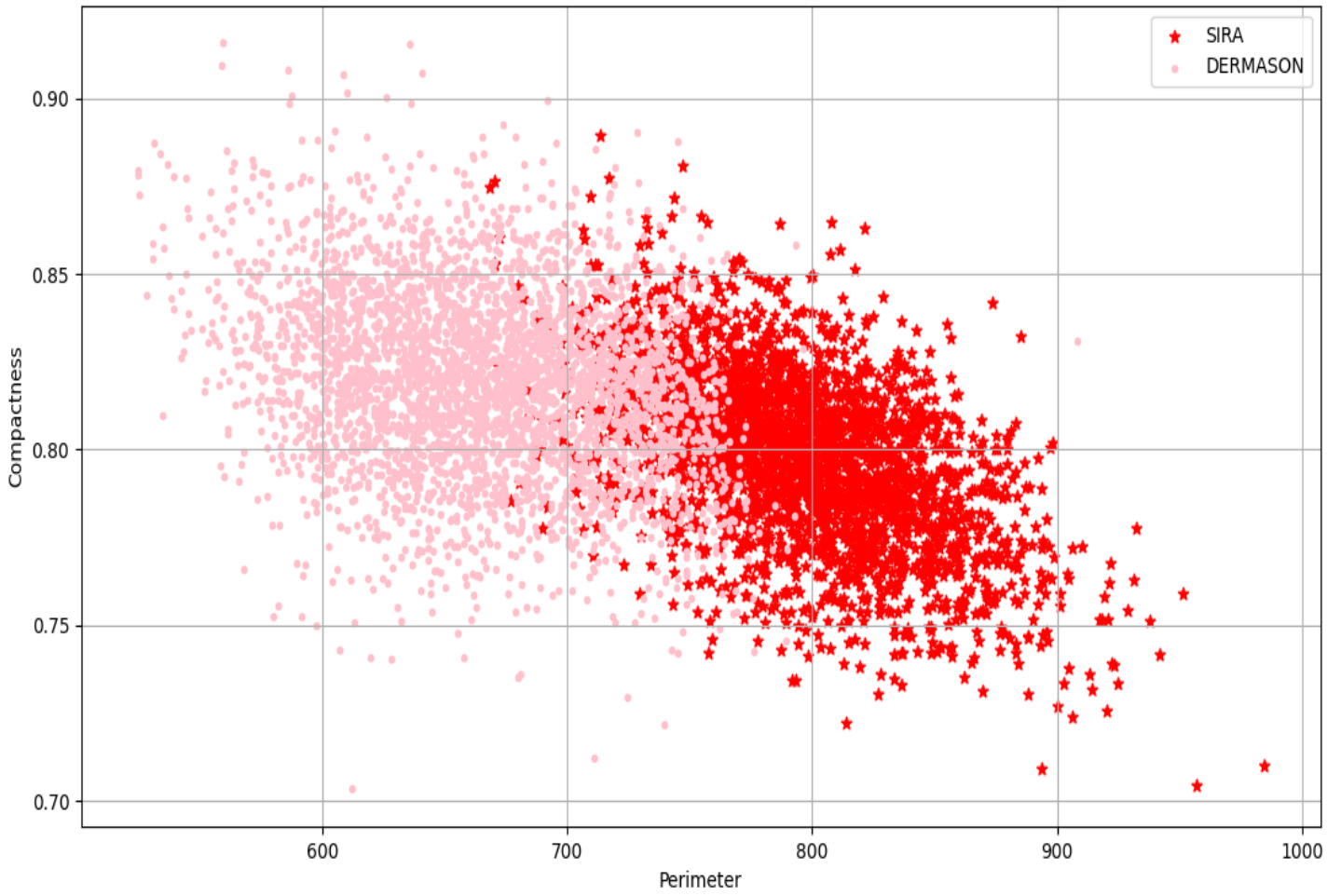
Şekil 19: Class Değişkenine Ait Seker ile Dermason Fasulyenin Perimeter ve Compactness Değişkenine Göre Dağılım Grafiği

Şekil 19 da Class değişkenine ait Seker ile Dermason fasulyenin Perimeter ve Compactness değişkenine göre dağılım grafiği verilmektedir.
Şeker fasulyenin Compactness değeri ve Perimeter değeri Dermason fasulyeden fazladır.



Şekil 20: Class Değişkenine Ait Bombay ile Horoz Fasulyenin Perimeter ve ShapeFactor3 Değişkenine Göre Dağılım Grafiği

Şekil 20 de Class değişkenine ait Bombay ile Horoz fasulyenin Perimeter ve ShapeFactor3 değişkenine göre dağılım grafiği verilmektedir.
Bombay fasulyenin Perimeter ve ShapeFactor3 değerleri Horoz fasulyeden fazladır. Bu grafikte çekirdek fonksiyonlarında en yüksek sonuçları elde ettiğimiz iki fasulye türünü karşılaştırmaktayız. İki fasulye türü arasındaki fark net bir şekilde gözlenmektedir.



Şekil 21: Class Değişkenine Ait Sıra ile Dermason Fasulyenin Perimeter ve Compactness Değişkenine Göre Dağılım Grafiği

Şekil 21 de Class değişkenine ait Sıra ile Dermason fasulyenin Perimeter ve Compactness değişkenine göre dağılım grafiği verilmektedir.
Sıra fasulyenin Perimeter değeri Dermason fasulyeden fazladır. Dermason fasulyenin de Compactness değeri Sıra fasulyeden fazla olduğu gözlenmektedir.

4 UYGULAMA

Bu kısım da Kuru Faulye veri setindeki tüm fasulye çeşitlerinin Perimeter ve Compactness özniteliğine göre çekirdek fonksiyonlarından elde edilen sonuçlar tablolaştırılmıştır. Bu sonuçlardan en yüksek model tahmini olanlar ve en düşük model tahmini olanlar bulunmuştur. Tablolarda elde edilen sonuçların bulunmasında SMOTE tekniği kullanılmıştır. Ayrıca en iyi sonuçlar aşağıdaki şekilde bulunmuştur.

- RBF çekirdek fonksiyonu ile kurulan modelimiz, en iyi sonucu $\gamma = 0.09$ ve $C = 10$ değerleri için vermektedir.
- Poly çekirdek fonksiyonu ile kurulan modelimiz, en iyi sonucu $\gamma = 1$ ve degree=2 değerleri için vermektedir.
- Sigmoid çekirdek fonksiyonu ile kurulan modelimiz, en iyi sonucu $\gamma = 0.01$ değeri için vermektedir.

Tablo 3: Perimeter İle Compactness Özniteliklerin Ve Fasulye Çeşidinin Şeker İle Dermason Olması Durumunda Elde Edilen Sonuçlar

ÇEKİRDEK FONKSİYONLARI	***PERİMETER-COMPACTNESS ÖZNİTELİĞİ VE ŞEKER FASULYE ÇEŞİDİ İÇİN ELDE EDİLEN SONUÇLAR***	***PERİMETER-COMPACTNESS ÖZNİTELİĞİ VE DERMASON FASULYE ÇEŞİDİ İÇİN ELDE EDİLEN SONUÇLAR***
DOĞRUSAL	**0.9471171502019831**	**0.9320602276900477**
RBF	**0.9753947851634227**	**0.9537275064267352**
POLYY	**0.8857877341167829**	**0.5872199779654792**
SİGMOİD	**0.9441792141020933**	**0.9305912596401028**

Tablo 3 de Perimeter ile Compactness özelliklerinin ve fasulye çeşidi Şeker ile Dermason olması halinde çekirdek fonksiyonlarının verdiği sonuçlar elde edilmiştir.

Fasulye çeşidi olan Şeker fasulye için en yüksek model tahmini %97,53 ile RBF çekirdek fonksiyonuyla bulunmuştur. Bu sonuca yakın olarak %94,71 ile Doğrusal çekirdek fonksiyonu gelmektedir. En düşük model tahmini ise %88,57 ile Poly çekirdek fonksiyondur.

Fasulye çeşidi olan Dermason fasulye için en yüksek model tahmini %95,37 ile RBF çekirdek fonksiyonuyla bulunmuştur. Bu sonuca yakın olarak %93,20 ile Doğrusal çekirdek fonksiyonu gelmektedir. En düşük model tahmini ise %58,72 ile Poly çekirdek fonksiyondur.

Tablo 4: Perimeter İle Compactness Özniteliklerin Ve Fasulye Çeşidinin Bombay İle Barbunya Olması Durumunda Elde Edilen Sonuçlar

ÇEKİRDEK FONKSİYONLARI	***PERİMETER-COMPACTNESS ÖZNİTELİĞİ VE BOMBAY FASULYE ÇEŞİDİ İÇİN ELDE EDİLEN SONUÇLAR***	***PERİMETER-COMPACTNESS ÖZNİTELİĞİ VE BARBUNYA FASULYE ÇEŞİDİ İÇİN ELDE EDİLEN SONUÇLAR***
DOĞRUSAL	**0.9996327579875137**	**0.8277634961439588**
RBF	**0.9996327579875137**	**0.9041498347410943**
POLYY	**0.9996327579875137**	**0.5200146896804995**
SİGMOİD	**0.9996327579875137**	**0.8277634961439588**

Tablo 4 de Perimeter ile Compactness özelliklerinin ve fasulye çeşidi Bombay ile Barbunya olması halinde çekirdek fonksiyonlarının verdiği sonuçlar elde edilmiştir.

Fasulye çeşidi olan Bombay fasulye için en yüksek model tahmini %99,96 ile tüm çekirdek fonksiyonları aynı sonucu bulmuştur. Bu durum tüm çekirdek fonksiyonlarının Bombay fasulyeyi en iyi şekilde ayırt edebildiğini gösterir.

Fasulye çeşidi olan Barbunya fasulye için en yüksek model tahmini %90,41 ile RBF çekirdek fonksiyonuyla bulunmuştur. Bu sonuca yakın olarak %82,77 ile Doğrusal ve Sigmoid çekirdek fonksiyonları gelmektedir. En düşük model tahmini ise %52,00 ile Poly çekirdek fonksiyondur.

Tablo 5: Perimeter İle Compactness Özniteliklerin Ve Fasulye Çeşidinin Sıra İle Calı Olması Durumunda Elde Edilen Sonuçlar

ÇEKİRDEK FONKSİYONLARI	***PERİMETER-COMPACTNESS ÖZNİTELİĞİ VE SIRA FASULYE ÇEŞİDİ İÇİN ELDE EDİLEN SONUÇLAR***	***PERİMETER-COMPACTNESS ÖZNİTELİĞİ VE CALI FASULYE ÇEŞİDİ İÇİN ELDE EDİLEN SONUÇLAR***
DOĞRUSAL	**0.5100991553433712**	**0.8057289753947852**
RBF	**0.9107601909658465**	**0.940506793977231**
POLYY	**0.8879911861917004**	**0.7921410209327947**
SİGMOİD	**0.4921042967315461**	**0.799118619170033**

Tablo 5 de Perimeter ile Compactness özelliklerinin ve fasulye çeşidi Sıra ile Calı olması halinde çekirdek fonksiyonlarının verdiği sonuçlar elde edilmiştir.

Fasulye çeşidi olan Sıra fasulye için en yüksek model tahmini %91,07 ile RBF çekirdek fonksiyonuyla bulunmuştur. Bu sonuca yakın olarak %88,79 ile Poly çekirdek fonksiyonu gelmektedir. En düşük model tahmini ise %49,21 ile Sigmoid çekirdek fonksiyondur.

Fasulye çeşidi olan Calı fasulye için en yüksek model tahmini %94,05 ile RBF çekirdek fonksiyonuyla bulunmuştur. Bu sonuca yakın olarak %80,57 ile Doğrusal çekirdek fonksiyonu gelmektedir. En düşük model tahmini ise %79,21 ile Poly çekirdek fonksiyondur.

Tablo 6: Perimeter İle Compactness Özniteliklerin Ve Fasulye Çeşidinin Horoz Olması Durumunda Elde Edilen Sonuçlar

ÇEKİRDEK FONKSİYONLARI	***PERİMETER-COMPACTNESS ÖZNİTELİĞİ VE HOROZ FASULYE ÇEŞİDİ İÇİN ELDE EDİLEN SONUÇLAR***
DOĞRUSAL	**0.9665809768637532**
RBF	**0.9761292691883952**
POLYY	**0.8688946015424165**
SİGMOİD	**0.964744766801322**

Tablo 6 da Perimeter ile Compactness özelliklerinin ve fasulye çeşidinin Horoz olması halinde çekirdek fonksiyonlarının verdiği sonuçlar elde edilmiştir.

Fasulye çeşidi olan Horoz fasulye için en yüksek model tahmini %97,61 ile RBF çekirdek fonksiyonuyla bulunmuştur. Bu sonuca yakın olarak %96,65 ile Doğrusal çekirdek fonksiyonu gelmektedir. En düşük model tahmini ise %86,88 ile Poly çekirdek fonksiyondur.

5 SONUÇ

Sonuç olarak; öncelikle işe veri setimizi tanıyarak ve özellikler arasındaki ilişkiyi yorumlayabilmek için korelasyon matrisi oluşturmakla başladık.
Sonra modelimizi oluştururken kullandığımız tekniklerin ne işe yaradığı ile ilgili bilgilendirmeler yapılmıştır. Bunlar; SMOTE tekniği, özellik ölçeklendirme, Random Forest algoritması, train ve test ayrımı, SVM ve çekirdek fonksiyonlarıdır.
SMOTE tekniği sayesinde veri setimizdeki veri dengesizliğini gidermiş olduk (Şekil 8-9).
Özellik ölçeklendirme ile özelliklerin ortalama değerini 0 ve standart sapmasını 1 olacak şekilde standart normal dağılıma uygun hale getirdik. Böylece özelliklerin aynı ölçekte olduğundan emin olduk.
Random Forest algoritması sayesinde veri setimizdeki 16 öznitelik arasından önem düzeyine göre bir sıralama elde etmiş olduk (Şekil 10). Bu algoritma sayesinde daha az sayıda ama en önemli özelliklerle ilgilenmiş oluruz. Bu yüzden hem zamandan tasarruf etmiş oluruz hem de iş yükümüz hafiflemiş olur.
Train ve test ayırımı ile kurduğumuz modelin başarısını test ederiz.
Artık kurduğumuz model ile çekirdek fonksiyonlarının karşılaştırmasını yapabilir hale gelmiş oluruz (Tablo 2). Bu sonuçlardan SMOTE tekniğinin model çalışmasında ki olumlu etkisi görülmüştür. Ayrıca en yüksek model doğruluğunu RBF çekirdek fonksiyonuyla elde edildiği gözlemlenmiştir.
Daha sonra veri setinde Class özniteliğine ait kuru fasulye çeşitlerinin Perimeter ve Compactness değişkenine göre dağılım ve çizgi grafiği oluşturulmuştur (Şekil 15-16). Class değişkenine ait verilerin ayrı ayrı Perimeter ve Compactness değişkenine göre dağılımını veren grafikler elde edilmiştir (Şekil 17-18). Ayrıca 2 farklı çeşit kuru fasulyenin 2 farklı özniteliğe göre dağılım grafikleri de elde edilmiştir.
Son olarak bu işlemler doğrultusunda farklı uygulamalar yapılmıştır. Bu uygulamalar sonucunda en yüksek model tahminini %99,96 ile tüm çekirdek fonksiyonları için Bombay fasulye çeşidi vermiştir.

6 KAYNAKÇA

(1) Bu Alanda Yapılmış Çalışmanın Alındığı Kaynak link olarak verilmiştir.
(2) Dry Bean Dataset. (2020). UCI Machine Learning Repository. UCI
(3) Bulut, F. (2016). Sınıflandırıcı Topluluklarının Dengesiz Veri Kümeleri Üzerindeki Performans Analizleri Faruk BULUT. Bilişim Teknolojileri Dergisi, 9(2), 153.
(4) Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence
(5) Qi, Y., Random forest for bioinformatics , (2013-20-4), 17 p.
(6) Akman, M., 2010, Veri madenciliğine genel bakış ve random forests yönteminin incelenmesi: sağlık alanında bir uygulama, Yüksek Lisans Tezi, Ankara Üniversitesi, Sağlık Bilimler Enstitüsü, 82 s.
(7) Cutler, A., Cutler, D. R. and Stevens, J. R., 2012, Ensemble machine learning, Springer, New York, 329 p.
(8) Cutler, A., Cutler, D. R. and Stevens, J. R., Tree-based methods, (2013-25-4),21 p.