

T.C.
VAN YÜZÜNCÜ YIL ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
İSTATİSTİK ANABİLİM DALI

**DESTEK VEKTÖR MAKİNESİ VE ÇEKİRDEK FONKSİYONLARI
ÜZERİNE BİR ÇALIŞMA**

YÜKSEK LİSANS TEZİ

Eyyup ADAR
Danışman: Dr. Öğr. Üyesi Sanem ŞEHRİBANOĞLU

VAN – 2023

T.C.
VAN YÜZÜNCÜ YIL ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
İSTATİSTİK ANABİLİM DALI

**DESTEK VEKTÖR MAKİNESİ VE ÇEKİRDEK FONKSİYONLARI
ÜZERİNE BİR ÇALIŞMA**

YÜKSEK LİSANS TEZİ

Eyyup ADAR

VAN – 2023

KABUL VE ONAY SAYFASI

Van Yüzüncü Yıl Üniversitesi Fen Bilimleri Enstitüsü, İstatistik Anabilim Dalı'nda Dr. Öğr. Üyesi Sanem ŞEHRİBANOĞLU danışmanlığında, Eyyup ADAR tarafından sunulan “Destek Vektör Makinesi ve Çekirdek Fonksiyonları Üzerine Bir Çalışma” başlıklı bu çalışma Lisansüstü Eğitim ve Öğretim Yönetmeliği'nin ilgili hükümleri gereğince 24/02/2023 tarihinde aşağıdaki jüri tarafından oy birliği / oy çokluğu ile başarılı bulunmuş ve yüksek lisans tezi olarak kabul edilmiştir.

Başkan: Doç Dr. Yılmaz KAYA

İmza:

Üye: Dr. Öğr Üyesi Sanem ŞEHRİBANOĞLU

İmza:

Üye: Dr. Öğr Üyesi Fatih KUTLU

İmza:

Fen Bilimleri Enstitüsü Yönetim Kurulu'nun /.... /..... tarih ve sayılı kararı ile onaylanmıştır.

İmza

.....

Enstitü Müdürü

ETİK BEYAN

Tez içindeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edilerek sunulduğunu, ayrıca tez yazım kurallarına uygun olarak hazırlanan bu çalışmada bana ait olmayan her türlü ifade ve bilginin kaynağına eksiksiz atıf yapıldığını bildiririm.

İmza
Eyyup ADAR



ÖZET

DESTEK VEKTÖR MAKİNESİ VE ÇEKİRDEK FONKSİYONLARI ÜZERİNE BİR ÇALIŞMA

ADAR, Eyyup

Yüksek Lisans Tezi, İstatistik Anabilim Dalı

Tez Danışmanı: Dr. Öğr. Üyesi Sanem ŞEHRİBANOĞLU

Şubat 2023, 95 sayfa

Bu tez çalışmasında on adet farklı veri seti üzerinde Destek vektör makineleri ve çekirdek fonksiyonları olan Lineer çekirdek, Polinom çekirdek, Radial Basis, Sigmoid çekirdek ve Pearson VII işlevsel çekirdek fonksiyonları ile sınıflandırma işlemi yapılarak performans ölçütleri açıklanmıştır.

İlk olarak sınıflandırma performansını etkileyen ceza parametresi olarak geçen C parametresi her bir çekirdek fonksiyonu için deneme yapılma yöntemi ile belirlenmiştir. Ardından doğruluk değeri, Kappa istatistik değeri ve ROC eğrisi altında kalan alan (AUC değeri) hesaplanarak sınıflar arasındaki uyum ve modelin başarı ölçütleri hesaplanmıştır. Tüm bu işlemlerden sonra çekirdek fonksiyonları arasındaki sınıflandırma başarısı değerlendirilmiş en yüksek başarı oranı olarak Lineer çekirdek ve Pearson VII işlevsel çekirdek fonksiyonu ile elde edildiği tespit edilmiştir.

Anahtar kelimeler: Destek vektör, Kappa, Lineer çekirdek, Pearson VII işlevsel çekirdek, Polinom çekirdek, Radial Basis çekirdek, ROC eğrisi, Sigmoid çekirdek

ABSTRACT

A STUDY ON SUPPORT VECTOR MACHINE AND KERNEL FUNCTIONS

ADAR, Eyyup

M. Sc. Thesis Department of Statistics

Supervisor: Asst. Prof. Dr. Sanem ŞEHRİBANOĞLU

February 2023, 95 pages

In this thesis, performance criteria are explained by classifying ten different data sets with Support vector machines and kernel functions such as Linear kernel, Polynomial kernel, Radial Basis, Sigmoid kernel and Pearson VII functional kernel functions.

Firstly, the C parameter, which is the penalty parameter that affects the classification performance, was determined by the trial method for each kernel function. Then, the accuracy value, Kappa statistical value and the area under the ROC curve (AUC value) were calculated, and the fit between the classes and the success criteria of the model were calculated. After all these processes, the classification success among the kernel functions was evaluated and it was determined that the highest success rate was obtained with the Linear kernel and Pearson VII functional kernel function.

Keywords: Kappa, Linear kernel, Pearson VII functional kernel, Polynomial kernel, Radial Basis kernel, ROC curve, Sigmoid kernel, Support vector



TEŞEKKÜR

Bu tez çalışmasında, her türlü ilgi ve yardımlarını esirgemeyen danışmanım Sayın Dr. Öğr. Üyesi Sanem ŞEHRİBANOĞLU'na teşekkürlerimi sunarım. Ayrıca tez jürimde yer alan desteklerini esirgemeyen Doç. Dr. Yılmaz KAYA, Sayın Dr. Öğr. Ü. Fatih KUTLU'ya teşekkürlerimi sunarım. Benden manevi desteklerini esirgemeyen değerli aileme, Derya ADAR, Şilan ADAR'a Yöneticim sayın Zehra BAKANGİL AKDOĞAN, dostum Ümit ÖZBEY, Murat ADAR ve Merve SAYHAN'a teşekkürlerimi sunarım.

2023

Eyyup ADAR



İÇİNDEKİLER

	Sayfa
ÖZET	i
ABSTRACT	iii
TEŞEKKÜR	v
İÇİNDEKİLER.....	vii
ÇİZELGELER LİSTESİ.....	ix
ŞEKİLLER LİSTESİ.....	xiii
SİMGELER VE KISALTMALAR	xvii
1. GİRİŞ.....	1
1.1 İstatistiksel Öğrenme Teorisi	3
2. KAYNAK BİLDİRİŞLERİ	5
3. MATERYAL VE YÖNTEM.....	9
3.1 Materyal	9
3.2. Yöntem.....	11
3.2.1 Ampirik Risk Minimizasyonu (ERM)	11
3.2.1.1 Büyüme Fonksiyonunun Yapısı	11
3.2.2 Yapısal Risk Minimizasyonu.....	13
3.2.3 Vapnik–Chervonenkis Teorisi (VC-Teorisi)	13
3.2.4 Destek Vektör Makinesi	14
3.2.5 Çok Sınıflı Destek Vektör Makineleri	16
3.2.5.1 Bire-Karşı-Bir.....	16
3.2.5.2 Bire-Karşı-Hepsi	17
3.2.5.3 DAG-SVM	17
3.2.6 Doğrusal Destek Vektör Makineleri	17
3.2.6.1 Doğrusal Ayrılabilen Verilerde İkili Sınıflandırma	17
3.2.6.2 Tamamen Doğrusal (Soft Marjin) Olarak Ayrılamayan Veriler İçin İkili Sınıflandırma.....	22
3.2.7 Doğrusal Olmayan Destek Vektör Makineleri	25
3.2.8 Çekirdek Fonksiyonları	26
3.2.8.1 Çekirdekleştirme	27
3.2.8.2 Doğrusal Çekirdek Fonksiyonu.....	28
3.2.8.3 Radial Basis Çekirdek Fonksiyonu (RBF)	28
3.2.8.4 Polinom Çekirdek Fonksiyonu	28

	Sayfa
3.2.8.5 Sigmoid Çekirdek Fonksiyonu	29
3.2.8.6 ANOVA Çekirdek Fonksiyonu	29
3.2.8.7 Pearson VII Evrensel İşlev Çekirdeği	29
3.2.8.8 String Çekirdek Fonksiyonu	30
3.2.9 PCA Analizi ile Destek Vektör Makinesi Formülasyonu ve Çekirdek Fonksiyonu	32
3.2.9.1 Çekirdek PCA	32
3.2.10 Özellik Seçimi (Future Selection)	34
3.2.10.1 Özellik Seçiminin Adımları	34
3.2.10.2 Alt Küme Üretimi	35
3.2.10.3 Özellik Seçim Yöntemleri	35
3.2.11 Performans ve Tahmin Ölçütleri	36
3.2.11.1 Karışıklık Matrisi	36
3.2.11.2 Kappa İstatistiği	36
3.2.11.3 ROC Eğrisi	37
4. BULGULAR	39
4.1 Göğüs Kanseri Veri Seti Analiz Sonuçları	39
4.2 Otizm Taraması Yetişkin Veri Seti Analiz Sonuçları	43
4.3 Raisin Veri Seti Analiz Sonuçları	48
4.4 Voice Veri Seti Analiz Sonuçları	53
4.5 Weather Veri Seti Analiz Sonuçları	58
4.6 Breast Cancer Veri Seti Analiz Sonuçları	62
4.7 Labor neg Veri Seti Analiz Sonuçları	67
4.8 Pima Diabets Veri Seti Analiz Sonuçları	71
4.9 German Credit Veri Seti Analiz Sonuçları	75
4.10 Inosphere Veri Seti Analiz Sonuçları	80
5. TARTIŞMA SONUÇ	89
KAYNAKLAR	91
ÖZ GEÇMİŞ	97

ÇİZELGELER LİSTESİ

Çizelge 4.1	Göğüs kanseri veri seti lineer çekirdek fonksiyonu analiz sonucu.....	39
Çizelge 4.2	Göğüs kanseri veri seti polinom çekirdek fonksiyonu analiz sonucu...	40
Çizelge 4.3	Göğüs kanseri veri seti RBF çekirdek fonksiyonu analiz sonucu.....	41
Çizelge 4.4	Göğüs kanseri veri seti sigmoid çekirdek fonksiyonu analiz sonucu ...	42
Çizelge 4.5	Göğüs kanseri veri seti pearson VII işlevsel çekirdek fonksiyonu analiz sonucu.....	43
Çizelge 4.6	Otizm taraması yetişkin veri seti lineer çekirdek fonksiyonu analiz sonucu	43
Çizelge 4.7	Otizm taraması yetişkin veri seti polinom çekirdek fonksiyonu analiz sonucu	44
Çizelge 4.8	Otizm taraması yetişkin veri seti RBF çekirdek fonksiyonu analiz sonucu	45
Çizelge 4.9	Otizm taraması yetişkin veri seti sigmoid çekirdek fonksiyonu analiz sonucu	46
Çizelge 4.10	Otizm taraması yetişkin veri seti pearson VII işlevsel çekirdek fonksiyonu analiz sonucu.....	47
Çizelge 4.11	Raisin veri seti lineer çekirdek fonksiyonu analiz sonucu.....	48
Çizelge 4.12	Raisin veri seti polinom çekirdek fonksiyonu analiz sonucu.....	49
Çizelge 4.13	Raisin veri seti RBF çekirdek fonksiyonu analiz sonucu	50
Çizelge 4.14	Raisin veri seti sigmoid çekirdek fonksiyonu analiz sonucu	51
Çizelge 4.15	Raisin veri seti pearson işlevsel VII çekirdek fonksiyonu analiz sonucu	52
Çizelge 4.16	Voice veri seti lineer çekirdek fonksiyonu analiz sonucu	53
Çizelge 4.17	Voice veri seti polinom çekirdek fonksiyonu analiz sonucu	54
Çizelge 4.18	Voice veri seti RBF çekirdek fonksiyonu analiz sonucu	55
Çizelge 4.19	Voice veri seti sigmoid çekirdek fonksiyonu analiz sonucu.....	56
Çizelge 4.20	Voice veri seti pearson VII işlevsel çekirdek fonksiyonu analiz sonucu	57
Çizelge 4.21	Weather veri seti lineer çekirdek fonksiyonu analiz sonucu	58
Çizelge 4.22	Weather veri seti polinom çekirdek fonksiyonu analiz sonucu	59

Çizelge 4.23	Weather veri seti RBF ve sigmoid çekirdek fonksiyonu analiz sonucu	60
Çizelge 4.24	Weather veri seti pearson VII İşlevsel çekirdek fonksiyonu analiz sonucu	61
Çizelge 4.25	Breast cancer veri seti lineer çekirdek fonksiyonu analiz sonucu	62
Çizelge 4.26	Breast cancer veri seti polinom çekirdek fonksiyonu analiz sonucu	63
Çizelge 4.27	Breast cancer veri seti RBF çekirdek fonksiyonu analiz sonucu	64
Çizelge 4.28	Breast cancer veri seti sigmoid çekirdek fonksiyonu analiz sonucu.....	65
Çizelge 4.29	Breast cancer veri seti pearson VII İşlevsel çekirdek fonksiyonu analiz sonucu.....	66
Çizelge 4.30	Labor neg veri seti lineer ve RBF çekirdek fonksiyonu analiz sonucu	67
Çizelge 4.31	Labor neg veri seti polinom çekirdek fonksiyonu analiz sonucu	68
Çizelge 4.32	Labor neg veri seti sigmoid çekirdek fonksiyonu analiz sonucu	69
Çizelge 4.33	Labor neg veri seti pearson VII İşlevsel çekirdek fonksiyonu analiz sonucu	70
Çizelge 4.34	Pima diabets veri seti lineer çekirdek fonksiyonu analiz sonucu	71
Çizelge 4.35	Pima diabets veri seti polinom çekirdek fonksiyonu analiz sonucu	72
Çizelge 4.36	Pima diabets veri seti RBF ve sigmoid çekirdek fonksiyonu analiz sonucu	73
Çizelge 4.37	Pima diabets veri seti pearson VII İşlevsel çekirdek fonksiyonu analiz sonucu.....	74
Çizelge 4.38	German credit veri seti lineer çekirdek fonksiyonu analiz sonucu	75
Çizelge 4.39	German credit veri seti polinom çekirdek fonksiyonu analiz sonucu...	76
Çizelge 4.40	German credit veri seti RBF çekirdek fonksiyonu analiz sonucu.....	77
Çizelge 4.41	German credit veri seti sigmoid çekirdek fonksiyonu analiz sonucu ...	78
Çizelge 4.42	German credit veri seti pearson VII İşlevsel çekirdek fonksiyonu analiz sonucu.....	79
Çizelge 4.43	Ionosphere veri seti lineer çekirdek fonksiyonu analiz sonucu	80
Çizelge 4.44	Ionosphere veri seti polinom çekirdek fonksiyonu analiz sonucu	81
Çizelge 4.45	Ionosphere veri seti RBF çekirdek fonksiyonu analiz sonucu.....	82
Çizelge 4.46	Ionosphere veri seti sigmoid çekirdek fonksiyonu analiz sonucu	83

Çizelge 4.47	Ionosphere veri seti pearson VII işlevsel çekirdek fonksiyonu analiz sonucu	84
Çizelge 4.48	Göğüs kanseri veri setinin çekirdek fonksiyonları ile elde edilen analiz sonuçları	85
Çizelge 4.49	Otizm taraması yetişkin veri setinin çekirdek fonksiyonları ile elde edilen analiz sonuçları.....	85
Çizelge 4.50	Raisin veri veri setinin çekirdek fonksiyonları ile elde edilen analiz sonuçları.....	86
Çizelge 4.51	Voice veri setinin çekirdek fonksiyonları ile elde edilen analiz sonuçları.....	86
Çizelge 4.52	Weather veri setinin çekirdek fonksiyonları ile elde edilen analiz sonuçları.....	86
Çizelge 4.53	Breast cancer veri setinin çekirdek fonksiyonları ile elde edilen analiz sonuçları	86
Çizelge 4.54	Labor neg veri setinin çekirdek fonksiyonları ile elde edilen analiz sonuçları.....	86
Çizelge 4.55	Pima diabets veri setinin çekirdek fonksiyonları ile elde edilen analiz sonuçları.....	87
Çizelge 4.56	German credit veri setinin çekirdek fonksiyonları ile elde edilen analiz sonuçları	87
Çizelge 4.57	Ionosphere veri setinin çekirdek fonksiyonları ile elde edilen analiz sonuçları.....	87
Çizelge 4.58	Veri setlerine ait özellikler ve en iyi performans gösteren çekirdek fonksiyonları	87



ŞEKİLLER LİSTESİ

Şekil 3.1	Verileri sınıflandırmak için birçok hiperdüzlem mevcut	15
Şekil 3.2	Maksimum ayrılmayı sağlayan hiperdüzlem	15
Şekil 3.3	Doğrusal olarak ayrılabilen iki sınıfa ait alt düzlem	18
Şekil 3.4	Doğrusal olarak ayrılamayan iki sınıfa ait alt düzlem.....	23
Şekil 3.5	Doğrusal ayrılamayan veriler	26
Şekil 3.6	Özellik seçimi akış şeması	35
Şekil 3.7	Kappa İstatistiğinin yorumlanmasına ilişkin değer aralıkları	37
Şekil 4.1	Göğüs kanseri veri seti lineer çekirdek fonksiyonu ile analiz edilen örneklere ait ROC eğrisi.....	39
Şekil 4.2	Göğüs kanseri veri seti polinom çekirdek fonksiyonu ile analiz edilen örneklere ait ROC eğrisi.....	40
Şekil 4.3	Göğüs kanseri veri seti RBF ile analiz edilen örneklere ait ROC eğrisi	41
Şekil 4.4	Göğüs kanseri veri seti sigmoid çekirdek fonksiyonu ile analiz edilen örneklere ait ROC eğrisi.....	42
Şekil 4.5	Göğüs kanseri veri seti pearson VII işlevsel çekirdek fonksiyonu ile analiz edilen örneklere ait ROC eğrisi.	43
Şekil 4.6	Otizm taraması yetişkin veri seti lineer çekirdek fonksiyonu ile analiz edilen örneklere ait ROC eğrisi.	44
Şekil 4.7	Otizm taraması yetişkin veri seti polinom çekirdek fonksiyonu ile analiz edilen örneklere ait ROC eğrisi.	45
Şekil 4.8	Otizm taraması yetişkin veri seti RBF çekirdek fonksiyonu ile analiz edilen örneklere ait ROC eğrisi.....	46
Şekil 4.9	Otizm taraması yetişkin veri seti sigmoid çekirdek fonksiyonu ile analiz edilen örneklere ait ROC eğrisi.	47
Şekil 4.10	Otizm taraması yetişkin veri seti pearson VII işlevsel çekirdek fonksiyonu ile analiz edilen örneklere ait ROC eğrisi.	48
Şekil 4.11	Raisin veri seti lineer çekirdek fonksiyonu ile analiz edilen örneklere ait ROC eğrisi.....	49

Şekil 4.12	Raisin veri seti polinom çekirdek fonksiyonu ile analiz edilen örneklerle ait ROC eğrisi.....	50
Şekil 4.13	Raisin veri seti RBF çekirdek fonksiyonu ile analiz edilen örneklerle ait ROC eğrisi.....	51
Şekil 4.14	Raisin veri seti sigmoid çekirdek fonksiyonu ile analiz edilen örneklerle ait ROC eğrisi.....	52
Şekil 4.15	Raisin veri seti pearson VII işlevsel çekirdek fonksiyonu ile analiz edilen örneklerle ait ROC eğrisi.....	53
Şekil 4.16	Voice veri seti lineer çekirdek fonksiyonu ile analiz edilen örneklerle ait ROC eğrisi.....	54
Şekil 4.17	Voice veri seti polinom çekirdek fonksiyonu ile analiz edilen örneklerle ait ROC eğrisi.....	55
Şekil 4.18	Voice veri seti RBF çekirdek fonksiyonu ile analiz edilen örneklerle ait ROC eğrisi.....	56
Şekil 4.19	Voice veri seti sigmoid çekirdek fonksiyonu ile analiz edilen örneklerle ait ROC eğrisi.....	57
Şekil 4.20	Voice veri seti pearson VII işlevsel çekirdek fonksiyonu ile analiz edilen örneklerle ait ROC eğrisi.....	58
Şekil 4.21	Weather veri seti lineer çekirdek fonksiyonu ile analiz edilen örneklerle ait ROC eğrisi.....	59
Şekil 4.22	Weather veri seti polinom çekirdek fonksiyonu ile analiz edilen örneklerle ait ROC eğrisi.....	60
Şekil 4.23	Weather veri seti RBF ve sigmoid çekirdek fonksiyonu ile analiz edilen örneklerle ait ROC eğrisi.....	61
Şekil 4.24	Weather veri seti pearson VII işlevsel çekirdek fonksiyonu ile analiz edilen örneklerle ait ROC eğrisi.....	62
Şekil 4.25	Breast cancer veri seti lineer çekirdek fonksiyonu ile analiz edilen örneklerle ait ROC eğrisi.....	63
Şekil 4.26	Breast cancer veri seti polinom çekirdek fonksiyonu ile analiz edilen örneklerle ait ROC eğrisi.....	64
Şekil 4.27	Breast cancer veri seti RBF çekirdek fonksiyonu ile analiz edilen örneklerle ait ROC eğrisi.....	65

Şekil 4.28	Breast cancer veri seti sigmoid çekirdek fonksiyonu ile analiz edilen örneklerle ait ROC eğrisi.....	66
Şekil 4.29	Breast cancer veri seti pearson VII işlevsel çekirdek fonksiyonu ile analiz edilen örneklerle ait ROC eğrisi.	67
Şekil 4.30	Labor neg veri seti lineer ve RBF çekirdek fonksiyonu ile analiz edilen örneklerle ait ROC eğrisi.....	68
Şekil 4.31	Labor neg veri seti polinom çekirdek fonksiyonu ile analiz edilen örneklerle ait ROC eğrisi.....	69
Şekil 4.32	Labor neg veri seti sigmoid çekirdek fonksiyonu ile analiz edilen örneklerle ait ROC eğrisi.....	70
Şekil 4.33	Labor neg veri seti pearson VII işlevsel çekirdek fonksiyonu ile analiz edilen örneklerle ait ROC eğrisi.....	71
Şekil 4.34	Pima diabets veri seti lineer çekirdek fonksiyonu ile analiz edilen örneklerle ait ROC eğrisi.....	72
Şekil 4.35	Pima diabets veri seti polinom çekirdek fonksiyonu ile analiz edilen örneklerle ait ROC eğrisi.....	73
Şekil 4.36	Pima diabets veri seti RBF ve sigmoid çekirdek fonksiyonu ile analiz edilen örneklerle ait ROC eğrisi.....	74
Şekil 4.37	Pima diabets veri seti pearson VII işlevsel çekirdek fonksiyonu ile analiz edilen örneklerle ait ROC eğrisi.	75
Şekil 4.38	German credit veri seti lineer çekirdek fonksiyonu ile analiz edilen örneklerle ait ROC eğrisi.....	76
Şekil 4.39	German credit veri seti polinom çekirdek fonksiyonu ile analiz edilen örneklerle ait ROC eğrisi.....	77
Şekil 4.40	German credit veri seti RBF çekirdek fonksiyonu ile analiz edilen örneklerle ait ROC eğrisi.....	78
Şekil 4.41	German credit veri seti sigmoid çekirdek fonksiyonu ile analiz edilen örneklerle ait ROC eğrisi.....	79
Şekil 4.42	German credit veri seti pearson VII işlevsel çekirdek fonksiyonu ile analiz edilen örneklerle ait ROC eğrisi.	80
Şekil 4.43	Ionosphere veri seti lineer çekirdek fonksiyonu ile analiz edilen örneklerle ait ROC eğrisi.....	81

Şekil 4.44	Ionosphere veri seti polinom çekirdek fonksiyonu ile analiz edilen örneklerle ait ROC eğrisi.....	82
Şekil 4.45	Ionosphere veri seti RBF çekirdek fonksiyonu ile analiz edilen örneklerle ait ROC eğrisi.....	83
Şekil 4.46	Ionosphere veri seti sigmoid çekirdek fonksiyonu ile analiz edilen örneklerle ait ROC eğrisi.....	84
Şekil 4.47	Ionosphere veri seti pearson VII işlevsel çekirdek fonksiyonu ile analiz edilen örneklerle ait ROC eğrisi.	85



SİMGELER VE KISALTMALAR

Bu çalışmada kullanılmış bazı simgeler ve kısaltmalar, açıklamalarıyla aşağıda sunulmuştur.

Simgeler

Açıklama

b	Destek vektör kat sayısı
b^*	Polinom çekirdek fonksiyonu parametresi
C	Hata düzenleme parametresi
C_{ov}	Kovaryans matrisi
D	Öz niteliklerin sayısı
d_1	H_1 'den orijine olan uzaklık
d_2	H_2 'den orijine olan uzaklık
f	Fonksiyon
F_i	i. adet fonksiyon dizimi
h	Hilbert Uzayı
H^+	α 'nın elde edilmesi sınıflandırıcı ve ve girdi değerlerine ait matris
H^{**}	Pearson VII işlevsel çekirdeğin tepe Yüksekliği
i	Veri setindeki her bir eleman
k	Çekirdek fonksiyonu
K	Kappa istatistiği
l	Etiketli örneklerin gözlem sayısı
L_p	α 'ya göre 1. dereceden kısmi türev
L_D	x_i girdi değerlerinin iç çarpımları
m	Hilbert uzayındaki öz vektör sayısı
N	Çoklu SVM' de sınıflandırıcı sayısı
$P(X,Y)$	Olasılık dağılımı
$\bar{P}$	Gözlemlenen uyumluluk oranı
$\bar{P}_e$	Şansa bağlı uyumluluk oranı
R^D	Düzlemdeki öz niteliklerin sayısı
R^n	Vektör sayısı
$R(\theta)$	Beklenen test hatası

Simgeler

Açıklama

$R_{emp}(\theta)$	Ampirik riskin ölçülen ortalama hatası
S	Destek vektör indisleri
u	x 'in bir alt dizini
V	Hilbert uzayında öz vektör
W	Hiperdüzlem normali
X	Girdi uzayı
x_i	Eğitim verisindeki i . Girdi örnekleri
x_i	Eğitim örnekleri arasında seçilen destek vektör matrisi sütununda yer alan girdi örnekler
x_s	Destek vektörü girdi değerleri
x'	Sonradan veri setine dahil olan nokta
Y	Çıktı uzayı
Y_i	Çıktı vektörüne ait -1 ve +1 değerleri
y_i	Çıktı vektörünün i . satırı
y_j	Çıktı vektörünün j . Sütunu
y_s	Destek vektörünün sınıf değeri
y'	x' noktasının sınıf değeri
q_i^n	Normalleştirme terimi
ξ_i	Soft margin'e ait gevşek değişken
μ_i	Eğitim verilerindeki hataların ortalaması
Φ	Özellik uzayı
$\Phi(x)$	Özellik vektörü
θ	f fonksiyonuna ait parametrelerin kümesi
Θ	Hipotez uzayının kümesi
λ	Hilbert uzayında öz değer
Ω	Boolean değerler
σ	RBF çekirdeğinin parametresi
α	Lagrange çarpanı
Σ	Alfabe içeren karakter kümesi
Σ^*	Σ üzerindeki boş olmayan kümeler
α_i	Eğitim örnekleri arasında seçilen destek vektör matrisi satır örnekleri

Simgeler**Açıklama** α_j

Eğitim örnekleri arasında seçilen destek vektör matrisi sütun örnekleri

 ω

Tepe noktasının yarı genişliği

Kısaltmalar**Açıklama****AUC.**

Area Under the Curve

ERM

Ampirik Risk Minimizasyonu

BT

Bilgisayarlı Tomografi

BCC

Breast Cancer Coimbra

COLT

Computational Learning Theory

DT

Decision Tree

FN

False Negative

FP

False Positive

KNN

K Yakın Komşu

LR

Logistic Regression

NB

Naive Bayes

PCA

Principal Component Analysis

RF

Random Forest

RBF

Radial Basis Fonksiyon

SVM

Support Vector Machine

SRM

Structural Risk Minimization

SLT

Statistical Learning Theory

TN

True Negative

TP

True Positive

YSA

Yapay Sinir Ağları



1. GİRİŞ

Destek Vektör Makineleri (Support Vector Machine-SVM) istatistiksel öğrenme teorisine dayalı bir gözetimli öğrenme algoritmasıdır (Geniou vd., 2000).

İstatistiksel Öğrenme Teorisi; verilen bir eğitim setinde çıktı verilerini tahmin edecek bir fonksiyonu bulmayı sağlar (Geniou vd., 2000).

Makine öğrenmesi; kısaca bir algoritması becerisi olarak tanımlanabilir, bu algoritma veri kümelerinden en uygun modelini oluşturmayı sağlamaktadır. Makine öğrenimi, istatistiksel yöntemleri kullanarak sınıflandırma ve tahminler yapmak üzere algoritmaları eğitmektedir.

Makine öğrenimi yöntemleri; literatürde, Denetimli öğrenme, denetimsiz öğrenme olarak ikiye ayrılmaktadır. Sonradan literatüre yarı- denetimli öğrenme, yöntemide eklenmiştir (Gökalp, 2022).

Eğitim verilerinde girdi değişkeni ile çıktı değişkeni arasındaki eşlemeyi öğrenen ve bu öğrenmeyi daha sonra modele dahil olacak olan ve etiketi bilinmeyen çıktıları tahmin etmek denetimli öğrenme yöntemi olarak geçmektedir. Denetimli öğrenmede veriler eğitim ve test verileri olarak ikiye ayrılmaktadır. Eğitim verilerinde anlamlı bir bilgi elde edilmeye çalışılmaktadır. Veriler sınıflandırılmadan önce sınıf bilgilerinin elde edileceği bir veri setine ihtiyaç duyulmaktadır. Akabinde bu verilerden elde edilen değerler test verilerinde kullanılır ve bu sayede öğrenme işlevi gerçekleştirilir. Denetimli öğrenme sınıflandırma ve regresyon yöntemlerinde tercih edilen bir öğrenme yöntemidir. Sınıflandırma işleminin yapıldığı SVM bir denetimli öğrenme yöntemidir (Bousquet vd., 2004; Metlek ve Kayaalp, 2020).

Sınıflandırma: Veri setinde yer alan verileri, daha önceki; belirtilmiş olan etiketli verilere uyacak şekilde ayırması işlemine denilmektedir.

Regresyon: Bir ya da birden çok fazla bağımsız değişken ile hedef değişken arasındaki ilişkinin modelini, matematiksel olarak ortaya çıkarmaya çalışan yöntemle denilmektedir (Metlek ve Kayaalp, 2020).

Denetimsiz öğrenme: Bu öğrenme yönteminde, veri seti etiketsiz verilerden oluşmaktadır. Model bu etiketsiz verilerden bulunan ilişkileri bulmaya çalışır. Eğitim etiketsiz verilerden oluşur. Model bu etiketsiz verilerden çıktı sonucunu anlamaya, tanımaya çalışır. Bundan dolayı sistemden istenilen sonuçlar bilinmez, denetimsiz

öğrenme yöntemi kümeleme yönteminden tercih edilen bir öğrenme yöntemidir (Gökalp, 2022).

Kümeleme: Veri setinde yer alan benzer özellikteki verileri bir araya getirme işlemine denilmektedir (Metlek ve Kayaalp, 2020).

Yarı-Denetimli öğrenme; denetimli ve denetimsiz öğrenme yöntemlerinin bir karmasıdır. Bu öğrenme yönteminde hem etiketli hem de etiketsiz veriler bulunur. Öğrenme yöntemi hem etiketli hem de etiketsiz verilerin birlikte kullanılmasıyla sağlanmaktadır (Gökalp, 2022).

Destek Vektör Makineleri (Support Vector Machine, SVM) ilk kez Vapnik ve çalışma arkadaşları tarafından COLT (Computational Learning Theory) konferansında önerilmiştir. 1995 yılında ise Corina Cortes ve Viladimir N. Vapnik tarafından günümüzde kullandığımız son şekli verilmiştir (Bilişik, 2011). SVM'ler birden fazla uygulama alanında en güçlü sınıflandırma kümeleme ve regresyon algoritmalarından biridir (Kim vd., 2003; Cervantes vd., 2019). SVM'in amacı farklı sınıflara ait veri noktalarını ayıran özellik uzayında karar sınırları oluşturmaktır (Naganna ve Deka, 2014). İlk zamanlarda sadece doğrusal verilerin sınıflandırılması için geliştirilen SVM daha sonra çok sınıflı doğrusal olmayan verilerin sınıflandırılması için de genelleştirilmiştir (Gürsoy ve Subaşı, 2008; Kavzoğlu ve Çölkesen, 2010).

SVM düzlemdeki verileri sınıflara ayırmak için verilerin arasına bir çizgi çizerek iki gruba ayırır. Çizilen sınır bu iki sınıfın verilerine olan uzaklığı, maksimum olacak şekilde belirlenmelidir. Sınırın çizilmesi iki sınıfa yakın ve birbirine paralel çizgiler çizilir. Bu çizgiler birbirlerine yaklaştırılarak sınıf çizgisi belirlenir. Burada SVM veriyi sınıflandırmak için en uygun modeli tahmin eder (Karakoyun ve Hacibeyoglu, 2018). SVM, girdi uzayındaki verileri, problemin doğrusal olarak ayrılabilir hale geldiği yüksek boyutlu bir özellik uzayına eşlemek için çekirdek fonksiyonlarını kullanır (Zhang vd., 2004; Wang vd., 2010). SVM'ler girdi uzayında orijinal veri kümeleri daha yüksek boyutlu, hatta sonsuz boyutlu öznitelik uzayına eşlemektedir. Böylece özellik uzayında sınıflandırma problemi daha basit hale gelir (Naganna ve Deka, 2014). SVM iki boyutlu uzayda doğrusal, üç boyutlu uzayda düzlemsel ve daha çok boyutlu uzayda ise hiperdüzlem olarak veriyi iki ve ikiden çok sınıfa ayırmaktadır (Kaynar vd., 2016). En iyi makine öğrenmesi algoritmaları arasında yer alan SVM genellikle yüz tanıma, parmak izi tanıma gibi nesneleri algıma problemlerinde, hatalı kart tespiti vb. birçok sınıflandırma

problemine uygulanır (Tong ve Koller, 2001; Pradhan, 2012). Problemler, ikili problemin çözüldüğü, genellikle ikinci dereceden programlar olan dışbükey optimizasyon problemleri olarak formüle edilir (Schölkopf ve Smola, 2002). Sınıflandırma problemlerinin çözümünde başarılı sonuçlar elde etmiş, genelleme performansı yüksek bir algoritma olan SMV'nin avantajı; sınıflandırma problemini çözerken bunu kareli optimizasyon problemine dönüştürerek çözmesidir. Bu durumunun sağladığı avantaj ise, problemin çözümüne ulaşılırken, öğrenme aşamasında işlem sayısını azaltmasıdır. Buda SVM' in diğer algoritmalara göre daha hızlı çözüme ulaşmasını sağlamaktadır. (Osowski, 2004; Siwekand ve Markiewicz, 2004). SVM' in bu özelliği, büyük hacimli veri setlerinde avantaj teşkil etmektedir. Optimizasyon temelli bir algoritma olan SVM sınıflandırma performansı, kullanışlılık ve hesaplamaların karmaşık olduğu durumlarda diğer algoritmalara göre daha başarılıdır. Ayrıca farklı veri setlerinde problemin çözümünde, SVM' in uygun çekirdek fonksiyonu ile birlikte uygulanması, önemli bir rol oynamaktadır (Nitze, 2012; Schulthess ve Asche, 2012).

1.1 İstatistiksel Öğrenme Teorisi

İstatistiksel Öğrenme Teorisi (Statistical Learning Theory – SLT) Vapnik tarafından geliştirilen, Ampirik Risk Minimizasyonu (Empirical Risk Minimizasyonu - ERM) tümevarım ilkesine dayanmaktadır. SLT, aynı zamanda Vapnik Chervonenkis (VC) teorisi olarak da bilinir (Kecman, 2001; Cherkassky ve Mulier, 2007; Abbasi vd., 2010). SLT'nin temel sorusu, seçilen fonksiyonun ne kadar iyi genelleştirdiği veya daha önce görülmemiş girdiler için çıktıyı ne kadar iyi tahmin ettiğidir. Aynı zamanda herhangi bir fonksiyonun ampirik ve beklenen riski arasındaki mesafeye ilişkin olasılıksal sınırlar sağlar (Evgeniou vd., 2000; Poggio vd., 2002; Brunato ve Battiti, 2005). Temel fikir, basit fonksiyon sınıfları kullanıldığında iyi öğrenme garantisinin elde edilebileceği (Luxburg ve Schölkopf, 2011), gelecekteki noktalarda iyi genellemeler yapan fonksiyonların en iyi nasıl oluşturulacağı problemini matematiksel olarak ele almaktadır (Bredensteiner ve Bennett, 1999).

Bu tez çalışmasında, mevcut fonksiyonu ne kadar iyi genelleştirdiğini gösteren ve bu fonksiyonun ampirik ve beklenen riski arasındaki mesafeye olasılıksal sınırlar sağlayan ve ampirik risk minimizasyonu tüme varım ilkesine dayalı olan istatistiksel

öğrenme teorisine değinilmiştir. Olası fonksiyonların bir modelini seçen ve ampirik riski en aza indiren ampirik risk minimizasyonuna aynı zamanda VC boyutunun sonlu olmasına değinilmiştir. Uygun bir VC boyutunun seçilmesi için yapısal risk minimizasyonu ve ampirik risk minimizasyonu ilkesine duyarlı ve tümevarımsal öğrenmeyi amaçlayan VC teorisi anlatılmıştır. Destek vektör makinesi, çoklu destek vektör makinesi ve ikili sınıflandırma için doğrusal ve doğrusal ayrılamayan destek vektör makineleri anlatılmıştır. Destek vektör makinelerinin model performansını ve sınıflandırma doğruluğunu etkileyen parametreler ve çekirdek fonksiyonları kavramları matematiksel formülleriyle birlikte açıklanmıştır. Özellik sayısının fazla olduğu veri kümelerinde kullanılacak öğrenme algoritması için faydalı olabilecek özellikleri seçerek veri boyutunu küçülten özellik seçimi kavramına değinilmiştir.

2. KAYNAK BİLDİRİŞLERİ

Ayhan ve Erdoğan (2014), SVM'in çekirdek fonksiyonları yardımıyla sınıflandırma başarısı ölçtükleri çalışmalarında; radyal tabanlı, polinomiyal, lineer ve sigmoid çekirdek fonksiyonlarına yer verilmiştir. 12 adet veri seti kullandıkları bu çalışmada, dördü bankacılık, üçü tıp, birer adet bilgisayar sistemleri, fizik, kimya biyoloji ve hukuk olmak üzere yedi farklı veri seti kullanmışlardır. Kullandıkları veri setleri ait özellikler ise, Australian credit verisi, 6 numerik 8 kategorik toplam 14 özellik ve 690 örnekten, Bank veri seti, 7 numerik, 9 kategorik olmak üzere 16 özellik ve 4521 örnekten, German credit verisi 20 kategorik özellik ve 1000 adet örnekten, Heart Disease verisi 7 numerik 6 kategorik toplam 13 özellikten ve 270 örnekten, Ionosphere 34 numerik ve 351 örnekten, Pima indian diabet, 8 numerik özellik ve 768 örnekten, Spambase verisi 57 numerik özellikten ve 4601 örnekten, Wisconsin breast cancer verisi, 10 numerik özellik ve 699 örnekten, Türkiye kredi verisi, 4 numerik 13 kategorik toplam 17 özellikten ve 167 örnekten oluşan 2 sınıflı veri setleridir. Glass verisi 9 numerik özellik ve 214 örnekten oluşan 6 sınıflı, Iris verisi, 4 numerik özellik ve 150 örnekten oluşan 3 sınıflı, Wine verisi, 13 numerik özellik ve 178 örnekten oluşan 3 sınıflı veri setidir. Bu veri setleri üzerinde yaptıkları çalışmaya göre; Australian credit (%88.52) German credit (%79.40), Pima Indian diabet (%78.13), ionosphere (%94.32), bank (%89.60), Türkiye kredi (%100), glass (%95.46) eğitim verileri için en iyi sınıflandırma başarısı radyal tabanlı çekirdek fonksiyonu ile elde edildiği tespit etmişlerdir. Lineer çekirdek fonksiyonu, heart disease (%87.40) ve spambase (%94.60) eğitim veri seti için en iyi sınıflandırma başarısını gösterdiği tespit edilmiştir. Lineer ve polinom çekirdek fonksiyonları Iris eğitim veri seti için %98 ile en iyi sınıflandırma başarısı elde etmişlerdir. Polinom ve sigmoid çekirdek fonksiyonları için en iyi başarı oranı wine eğitim veri seti için %97.66 ve %100 olarak elde etmişlerdir. Tüm bu sonuçlara göre eğitim veri setleri için radyal tabanlı çekirdek fonksiyonun en iyi sınıflandırma başarısı elde ettiklerini ifade etmişlerdir.

Toğaçar ve Ergen (2019), 700 adet tümör görüntünün analiz edilmesi için konvolüsyonel sinir ağı yöntemini kullanmışlardır. 82 hastadan elde edilen görüntülerin 24'ü iyi huylu göğüs kanseri; 58'i ise kötü huylu göğüs kanseridir. 700 adet veri kümesinin 350 adeti iyi huylu diğer 350 adet ise kötü huylu veri kümesinden

oluşmaktadır. Verilerin sınıflandırılması için SVM algoritmasından faydalanılmış ve sınıflandırma başarısı %93.43 olduğunu ifade etmişlerdir.

Sevli (2019), Göğüs kanseri üzerine yaptıkları çalışmada Wisconsin üniversite hastanesinde Dr. William H. Wolsberg tarafından toplanan ve araştırma için paylaşılan, göğüs kanseri bulgularını içeren 569 örnekten oluşan veri setini kullanmışlardır. Verileri %80 eğitim, %20 test verisi rastgele bölerek 5 farklı makine öğrenme algoritmasıyla sonuçlarını karşılaştırmışlardır. Kullandıkları makine öğrenme algoritmaları; SVM), Naive Bayes (NB), Rastgele Orman (RF), KNN) ve Logistic Regresyon (LR) algoritmalarıdır. SVM %93.86, NB %94.74, RF %96.49, KNN %91.23, LR %98.24 doğruluk oranı elde ettiklerini tespit etmişlerdir.

Ecemiş vd. (2019), On adet farklı veri seti üzerinde çalışma yapmışlardır. Kullandıkları algoritmalar, Destek Vektör SVM, Naive Bayes (NB), K-En Yakın Komşu (KNN), Rastgele Orman (RF) algoritmalarıdır. Kullandıkları veri setleri;

Avustralya yağmur ormanı verisi, 24 özellik ve 142193 adet nominal ve numerik örnekten oluşan kategorik bir veri setidir. Rahim kanseri verisi 36 özellik 858 satırdan oluşan, hastanın kanser olup olmadığı bilgisini bulundurmaktadır. Ağaç değerlendirme veri seti, 7 özellik ve 1725 örnekten oluşmaktadır. Kalp hastalıkları veri seti, 14 özellik ve 303 örnekten oluşan numerik verilerden oluşan kategorik bir veri setidir. Ev satış veri seti, 16 özellik ve 21613 örnekten oluşmaktadır. Numerik verilerden oluşan kategorik bir veri setidir. Parkinson veri seti, 24 özellik ve 195 örnekten oluşmaktadır. Kategorik ve numerik verilerden oluşmaktadır. Pulsar yıldızı veri seti, 1'i sınıf değeri olmak üzere 9 farklı özellikten oluşan numerik değerler içeren kategorik bir veri setidir. Abalone veri seti 9 özellik ve 4177 adet örnekten oluşmaktadır. Mantar veri seti, 23 özellik 8124 örnekten oluşan nominal veri içeren kategorik bir veri setidir. Avustralya yağmur ormanı veri seti doğruluk oranı SVM %99.85, NB %94.14, KNN %76.2, RF %99.93 elde etmişlerdir. Araç değerlendirme veri seti doğruluk oranları SVM %99.42, NB %83.53, KNN %93.51, RF %94.50 elde etmişlerdir. Rahim kanseri veri seti doğruluk oranları SVM %99.30, NB %95.22, KNN %99.53, RF %99.30 elde etmişlerdir. Kalp hastalıkları veri seti SVM %84.81, NB %82.82, KNN %96.41, RF %91.79 elde etmişlerdir. Parkinson veri seti SVM %87.17, NB %69.23, KNN %96.41, RF %91.79 elde etmişlerdir. Pulsar yıldızı veri seti, SVM %98, NB %94.49, KNN %97.14 RF %97.99 ve mantar veri seti için SVM %100, NB %95.82, KNN %100, RF %100 doğruluk oranları

elde etmişlerdir. Çalışma sonucunda elde ettikleri bulgular doğrultusunda SVM algoritmasının Rahim kanseri, Pulsar yıldızı, araç değerlendirme ve kalp hastalıkları veri setlerinde en yüksek doğruluk oranın elde ettiğini tespit etmişlerdir. Ayrıca bunun nedeninin C parametresi ve doğrusal ayrılamayan veriler için kullandıkları radyal tabanlı çekirdek fonksiyonunun etikisi olduğu tespit etmişlerdir. Avustralya yağmur verisi için en iyi doğruluk oranına RF sahip olduğu yine aynı sonucu SVM algortimasının da elde ettiğini bulmuşlardır.

Özcan (2020), ise Breast cancer coimbra (BCC) veri seti ile göğüs kanseri üzerine yaptığı çalışmada SVM, K- En Yakın Komşu (KNN), Naive Bayes (NB) ve Karar Ağaçları (DT) algoritmasını kullanmıştır. Verileri min-max noarmalizasyon ön işleme tabi tutmuşlardır. Verileri ön işleme tabi tutmadan SVM ile %68.97, KNN %51.72, NB %68.97 ve DT için %62.07 başarı oranı elde ettiklerini tespit etmişlerdir. Verileri ön işleme tabi tuttukten sonra ise SVM ile %72.41, KNN %72.41, NB %65.52 ve DT %51.72 ile sınıflandırma başarı elde ettiklerini tespit etmişlerdir.

Tarihi vd. (2020), Firmaların başarısızlık durumlarına göre SVM yardımıyla sınıflandırmayı amaçlamışlardır. Başarısızlık durumu için iki farklı değişken belirleyerek; üç yıl üst üste zarar eden firmalar başarısız kabul edilmiş daha sonra iki yıl üst üste zarar eden firmalar başarısız kabul edilmiştir. Çalışma sonucunda; düşük teknoloji sınıfında olan işletmelerde üç yıl üst üste zarar eden firmalar için en iyi sınıflandırma performansı %92.59 ile radyal tabanlı çekirdek fonksiyonu olmuştur. İki yıl üst üste zarar eden firmalar için ise %92.59 ile polinom çekirdek fonksiyonu en iyi sınıflandırma başarısı gösterdiği tespit etmişlerdir. Orta- düşük teknoloji sınıfında olan işletmeler için üç yıl üst üste zarar eden firmalar için en iyi sınıflandırma performansı %87,80 ile sigmoid çekirdek fonksiyonu olduğu gözlemlenmiştir. İki yıl üst üste zarar eden firmalar için ise sigmoid ve polinom çekirdek fonksiyonu %85.37 ile en iyi sınıflandırma başarısı gösterdiği tespit etmişlerdir. Orta-ileri teknoloji sınıfında olan işletmeler tüm kernel fonksiyonlarının en iyi sınıflandırma başarısı gösterdiğini tespit etmişlerdir

Şimşek vd. (2021), doğal akarsularda debiyi ölçmek ve su seviyesini kontrol etmek için yaptıkları çalışmada debi katsayısının belirlenmesi üzerine SVM sınıflandırma başarısını ölçmek için gauss çekirdek fonksiyonu, polinom çekirdek fonksiyonu ve lineer çekirdek fonksiyonlarını kullanmışlardır. Bu çalışma sonucunda gauss çekirdek

fonksiyonunun diğer çekirdek fonksiyonlarına göre %93.91 ile daha başarılı olduğu ifade etmişlerdir.

Özdet ve İçer (2022), yaptıkları çalışmada, bilgisayarlı tomografi (BT) görüntülerini kullanarak tümör ve nodüllerinin tespit edilmesi, görüntülerden çıkarılan özellikleri farklı sınıflandırma algoritmaları ile sınıflandırmayı amaçlamışlardır. 26 farklı görüntü üzerinde yaptıkları çalışmada, SVM, Karar ağaçları, K-Yakın komşuluk (KNN) ve Diskriminant analiz ile sınıflandırılarak sonuçları karşılaştırmışlardır. Karar ağaçları %97, SVM %96.6, KNN %93.6, Diskriminant analizi %97 sınıflandırma sonuçlarını elde ettiklerini tespit etmişlerdir.

Bayrak vd. (2022), Göğüs kanseri sınıflandırması için kaggle platformunda elde ettikleri iki adet göğüs kanseri verisi üzerinde çalışmışlardır. Birinci veri seti 30 adet özellik ve 569 örnekten, ikinci veri seti 5 özellik ve 569 örnekten oluşmaktadır. Veri seti Wisconsin üniversitesinde elde edilmiştir. İki farklı göğüs kanseri veri setinde kategorik verilerin numerik verilere dönüştürülmesi, gereksiz verilerin atılması ve normalizasyon gibi ön işlemlerden sonra, K-En Yakın Komşu (KNN), Destek Vektör Makineleri (SVM), Karar Ağacı (DT), Naive Bayes (NB) ve Yapay Sinir Ağları (YSA) algoritmalarını kullanmışlardır. SVM için C parametresinin değerini 1 ve kullanılan çekirdek fonksiyonları olan Gamma ve Radyal basis için parametre değerlerini otomatik almışlardır. Birinci veri seti için elde ettikleri doğruluk oranları, KNN %94.73, SVM %97.36, DT %86.84, NB %92.10, YSA %98.24 ikinci veri seti için ise KNN %90.35, SVM %92.10, DT %85.08, NB %89.47, YSA %93.35 doğruluk oranlarını elde etmişlerdir.

3. MATERYAL VE YÖNTEM

3.1 Materyal

Bu tez çalışmasının uygulama bölümünde, on adet veri seti üzerinde çalışılmıştır. **Birinci veri seti:** Wisconsin Üniversitesinde Dr. William H. Wolsberg tarafından toplanan ve herkese açık paylaşılan, 569 örnekten oluşan göğüs kanseri veri seti üzerinde çalışılmıştır. Bu çalışmada kullanılan veri seti kaggle platformunda elde edilmiştir (<https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data?select=data.csv>). Veri seti 30 tanımlayıcı ve 1 teşhis özelliği olmak üzere toplam 31 özellik içermektedir. 30 tanımlayıcı özellik göğüste görülen kitle resimlerinin sayısallaştırılmasıyla toplanan, dokunun çapı, şekli ve pürüzsüzlüğü, yüzey alanı gibi verilerden oluşmaktadır. Tanımlayıcı özelliklerden 10 tanesi tümör hücresinin çekirdeği üzerinde direkt ölçümlenmiş, geri kalan 20 adet tanımlayıcı özellik ise bu tanımlayıcı özelliklere bağlı olarak hesaplanmış sayısal değerleri ifade etmektedir. Direkt ölçümlenen 10 özellik ise; yarıçap, doku, alan, pürüzsüzlük, yoğunluk, içbükeylik, içbükey nokta sayısı, simetri, fraktal ve boyuttur. Diğer özellikler ise, standart hata, ortalama, en kötü ve en büyük gibi değerlerdir. Teşhis sınıfı bu özelliklere bağlı olarak iyi huylu (benign) ve kötü huylu (malignant) etiketlerle ifade edilmektedir. 569 örneğin sınıf etiketi dağılımı ise; 212 kötü huylu, 357 iyi huylu şeklindedir.

İkinci veri seti: UCI Machine Learning Repository sitesinde elde edilmiştir (<https://archive.ics.uci.edu/ml/datasets/Autism+Screening+Adult>). Otizm taraması yetişkin veri seti; bir tanımlayıcı özellik olmak üzere toplamda 21 özellik ve 704 örnekten oluşmaktadır. 2 Adet özellik numerik 19 özellik ise string özelliklerden oluşmaktadır.

Üçüncü veri seti: Raisin veri seti kaggle platformunda elde edilmiştir (<https://www.kaggle.com/datasets/muratkokludataset/raisin-dataset>). Türkiye’de yetişen iki farklı üzüm çeşidi (keçimen ve besni) eşit olmak üzere 900 adet örnekten oluşmaktadır. Bir tanımlayıcı örnek olmak üzere 8 özellikten oluşan veri seti numerik örnekler içermektedir.

Dördüncü veri seti: UCI Machine Learning Repository sitesinde elde edilmiştir (UCI Machine Learning Repository). Voice veri seti 1 tanımlayıcı özellik toplamda 21

özellikten oluşan veri seti 3168 örnekten oluşmaktadır. Numerik özelliklerden içermektedir.

Beşinci veri seti: weka programında elde edilmiştir. Weather veri seti 1 tanımlayıcı özellik olmak üzere toplam 5 özellik ve 14 örnekten oluşmaktadır. String ve numerik veriler yer almaktadır.

Altıncı veri seti: weka programında elde edilmiştir. Breast cancer veri seti 1 tanımlayıcı özellik olmak üzere toplam 9 özellik ve 768 örnekten oluşmaktadır. Numerik ve string verilerden oluşmaktadır.

Yedinci veri seti: weka programında elde edilmiştir. Labor neg veri seti 1 tanımlayıcı olmak üzere toplam 17 özellik ve 57 örnekten oluşmaktadır. Numerik ve string verilerden yer almaktadır.

Sekizinci veri seti: weka programında elde edilmiştir. Pima diabets veri seti 1 tanımlayıcı olmak üzere toplamda 9 özellik ve 768 örnekten oluşmaktadır. Verilerin tamamı numeriktir.

Dokuzuncu veri seti: weka programında elde edilmiştir. German credit veri seti 1 tanımlayıcı olmak üzere toplam 21 özellik ve 1000 adet örnekten oluşmaktadır. Nominal ve numerik verilerden oluşmaktadır.

Onuncu veri seti: weka programında elde edilmiştir. Ionosphere veri seti 1 tanımlayıcı olmak üzere toplam 35 özellik ve 351 örnekten oluşmaktadır. Numerik veriler yer almaktadır.

Bu çalışmada kullanılan çekirdek fonksiyonları; Lineer çekirdek, Polinom çekirdek, Radial basis çekirdek, Sigmoid çekirdek ve Pearson VII işlevsel çekirdek fonksiyonlarıdır. Verilerin sınıflandırılması için makine öğrenmesinde sıkça kullanılan Weka (Waikato Environment for Knowledge Analysis) paket programının 3.8.5 versiyonu kullanılmıştır. Göğüs kanseri verisi, cross validation yöntemiyle 10 parçaya bölünerek SVM algoritmasının LIBSVM kütüphanesi kullanılarak sınıflandırılmıştır.

3.2. Yöntem

3.2.1 Ampirik Risk Minimizasyonu (ERM)

Ampirik risk minimizasyonu modellerin gerçek çıktı değerinden farklı bir çıktı değeri elde etmesi işlemidir. Bu, hata veya risk olarak tanımlanabilir (Korkmaz, 2020).

ERM basit fakat verimli algoritmalarından biridir. ERM olası fonksiyonların bir modelini seçer ve bu modeldeki ampirik riski en aza indirmeyi amaçlar (Bousquet vd., 2004). Kısacası ERM örnekler arasındaki hata oranını azaltmaya çalışır (Erdal, 2015). ERM tümevarım ilkesine dayanır. ERM ilkesinin tutarlı olması için gerekli ve yeterli şart VC boyutunun (h) sonlu olmasıdır. Tutarlılık teorisi; gözlemlerin sayısı arttıkça, üzerinde çalışılan öğrenme algoritması kullanılarak elde edilen mevcut çözümlerin mümkün olan en iyi çözüme yakınsaması için gerek ve yeterli şartları tanımlamaktadır. Öğrenme algoritmasında, algoritma çıktı olarak gördüğü tüm örnekleri doğru sınıflandırıyorsa tutarlıdır. “Eş. 1.0” herhangi bir olasılık ölçüsünde ERM’nin tutarlılığı için gerek ve yeter şartları sağlamaktadır.

$$\lim_{l \rightarrow \infty} \frac{G^{\wedge}(l)}{l} = 0 \quad (3.1)$$

$G^{\wedge}(l)$ büyüme fonksiyonunu ifade etmektedir. VC boyutu $G^{\wedge}(l)$ büyüme fonksiyonunun özelliğinden gelmektedir. (Brunato ve Battiti, 2005).

3.2.1.1 Büyüme Fonksiyonunun Yapısı

Teorem: herhangi bir büyüme fonksiyonu ya

$$G^{\wedge}(h) = h \ln 2, G^{\wedge}(h + 1) \neq (h + 1) \ln 2, h\text{'nin bir tamsayıdır.}$$

$$G^{\wedge}(l) < h(\ln \frac{1}{h} + 1) \text{ eşitsizliği tarafından sonlandırılır veya}$$

$G^{\wedge}(l) = l \ln 2$ eşitliğini sağlar. Farklı bir ifadeyle büyüme fonksiyonu ya doğrusal bir fonksiyon olacaktır ya da logaritmik bir fonksiyon tarafından sonlandırılacaktır. VC boyutunun sonlu olması olasılık ölçüsünden bağımsız olarak ERM yönteminin tutarlılığı için gerek ve yeter şartı sağlamaktadır (Tolun, 2008).

ℓ denetimli öğrenmede etiketli örneklerin gözlem sayısını belirtiyor olsun. Sınıflandırma probleminde örnekler $(x_i, y_i), i = 1, \dots, \ell$ çiftinden oluşur. Burada $x_i \in \mathbb{R}^n$ vektör ifade eder $y_i \in \{-1, +1\}$ süpervizör tarafından atanan etiketlerdir.

$P(x, y)$ örneklerin alındığı bilinmeyen olasılık dağılımını ifade ediyor olsun. Burada öğrenmenin görevi $f(x, \theta)$ fonksiyonunun parametrelerini belirleyip $x_i \rightarrow y_i$ eşlemesini öğrenmektir. $f(x, \theta)$ fonksiyonu hipotezi ifade eder. $\{f(x, \theta), \theta \in \Theta\}$ kümesi ise hipotez uzayını ifade eder ve h ile gösterilir. θ ise parametrelerin kümesini gösterir. Eğitilmiş makine etiketli örneklerle dayalı olan $\theta \in \Theta$ parametresinin seçimi ile belirlenir. Eğitilmiş bir makinenin beklenen test hatası ya da beklenen riski aşağıdaki şekilde ifade edilir.

$$R(\theta) = \int \|y - f(x, \theta)\| dP(x, y) \quad (3.2)$$

Ampirik risk ise eğitim setinde ölçülen ortalama hata oranıdır ve “Eş. 3.3” de gösterilmiştir.

$$R_{emp}(\theta) = \frac{1}{\ell} \sum_{i=1}^{\ell} \|y_i - f(x_i, \theta)\| \quad (3.3)$$

“Eş. 1.1” de riski en aza indiren $f(x, \theta^*)$ fonksiyonuna, “Eş. 1.2” deki ampirik riski en aza indiren $f(x, \hat{\theta})$ fonksiyonu ile yaklaşılr. ERM ilkesinin işleyiş mantığı; eğer R_{emp} olasılıkta R 'ye yakınsarsa minimum R_{emp} minimum R 'ye yakınsar. Bu durum sağlanmazsa ERM ilkesinin tutarlı olmadığı ifade edilir. ERM ilkesinin tutarlılığı için gerek ve yeter şart, h hipotez uzayının VC-boyutunun sonluluğudur. h 'nin (hipotez uzayı) VC-boyutuna değinecek olursak $f(x, \theta)$ fonksiyonları tarafından tüm olası şekillerde iki sınıfa ayrılabilen en fazla sayıda örnekten oluşmaktadır. VC-boyutu, hipotez uzayının karmaşıklığını ve tanımlayıcı gücünü ölçer ve çoğunlukla $f(x, \theta)$ modelinin serbest parametrelerinin sayısı ile orantılıdır (Martin ve ark., 2011).

3.2.2 Yapısal Risk Minimizasyonu

Vapnik tarafından geliştirilen Yapısal Risk Minimizasyonu (Structural Risk Minimization - SRM), uygun bir VC boyutu seçme problemini çözmeye çalışır (Trafalis ve Ince, 2000). SRM gerçek risk üzerindeki sınırı en aza indiren fonksiyonların alt kümesini bulmaya çalışır. Bu her alt küme için bir tane olacak şekilde bir dizi makineyi eğiterek yapılabilir. Burada belirli bir alt küme için eğitimin amacı deneysel riski en aza indirmektir (Burges, 1998a; Jakkula, 2006). Öğrenmenin amacı mevcut veri dağılımı üzerinde en iyi genelleme performansı sunan modeli bulmaktır. X girdi uzayı ve Y 'de çıktı uzayı olsun. $(X \times Y) \in X \times Y$ çiftlerinin bilinmeyen olasılık dağılımı $P(X, Y)$ olan rastgele değişkenler olduğunu varsayalım. Amacımız beklenen riski en aza indiren bir $f: X \rightarrow Y$ tahmin edicisini bulmaktır:

$$P(f(X) \neq Y) = \mathbb{E}_{(X,Y) \sim P_{XY}} [\delta(f(X) \neq Y)] \quad (3.4)$$

Pratikte $P(X,Y)$ 'den bağımsız olarak çizilmiş yalnızca n çift (X_i, Y_i) eğitim örneğine sahibiz. Ampirik riski en aza indirerek f fonksiyonu seçmek genelde aşırı uyuma (overfitting, modelin eğitim verilerinin her detayını öğrenip sonradan modele dahil olan verilere uygulamasıyla sınıflandırmanın olumsuz etkilenmesine) yol açar. Buradaki sorunu çözmek için yapısal risk minimizasyonu fikri $F_1, F_2, \dots$ sonsuz bir dizinini kullanmaktır. Burada her bir F_i bir dizi fonksiyondur. Modelin kapasitesi için bir ceza ile her modeldeki ampirik risk en aza indirilir:

$$f_n := \arg \min_{f \in F_i, i \in \mathbb{N}} \frac{1}{n} \sum_{j=1}^n \delta(f(X_j) \neq Y_j) + \text{capacity}(F_i, n) \quad (3.5)$$

Burada kapasite (F_i, n) , verilen eğitim seti bağlamında modeli F_i 'nin karmaşıklığını ölçer (Martin ve ark., 2011).

3.2.3 Vapnik–Chervonenkis Teorisi (VC-Teorisi)

VC-teorisi ERM ilkesine dayalı tümevarımsal öğrenmeyi amaçlayan önemli kavramları ve matematiksel sonuçları tanıtır. VC-teorisi 1950'ler ve 1960'larda

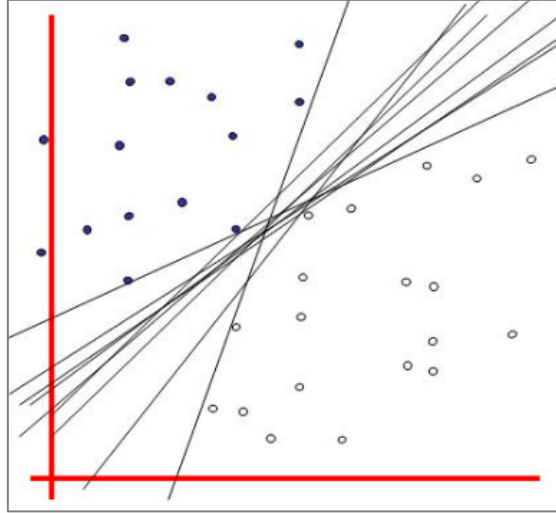
fizyologlar ve sinirbilimciler tarafından geliştirilen basit örüntü tanıma algoritmalarının teorik olarak daha iyi anlaşılmasını sağlamak amacıyla geliştirilmiştir (Cherkassky ve Mulier, 2007).

VC-teorisi, iyi tanımlanmış formülasyon ve kendi kendine tutarlı matematiksel teori ile ilgili temel kavram ve ilkeleri birleştirir. Ayrıca VC- teorisi; istatistik, sinir ağları, bulanık sistemler, sinyal işleme vb. alanlarda geliştirilen çeşitli öğrenme yöntemlerinin daha iyi anlaşılması için kullanılabilir (Widodo ve Yang, 2007).

3.2.4 Destek Vektör Makinesi

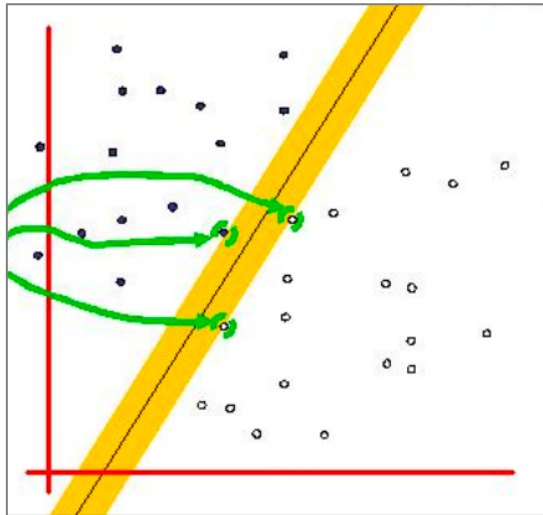
Destek vektör makinesi yapısal risk minimizasyonunu bünyesinde barındırarak, öğrenme modelinin beklenen hatasını en aza indirmektedir. Ayrıca aşırı uyum sorununu azaltarak daha iyi bir genelleme yapmaktadır. Bir algoritmanın eğitim setinde yer alan verileri doğru bir şekilde sınıflandırma yeteneği onun genelleme başarısı olarak tanımlanır (Jakkula, 2006; Widodo ve Yang, 2007). SVM modeli temel olarak girdi değerleri ile çıktı değerleri üzerine bir öğrenme işlevi oluşturur (Babaei vd., 2019). SVM doğru model formunun bilinmediği durumlarda doğru olabilecek modeller arasında optimal tahminleyici modelin seçilmesi esasına dayanmaktadır. Başka bir ifadeyle iki sınıfı birbirinden en uygun şekilde ayıran hiperdüzlemin elde edilmesi esasına dayanmaktadır (Kavzaoğlu ve Çölkesen, 2010). SVM’de sınıfları ayırmak için birçok hiperdüzlem yerleştirilebilir, ancak diğer hiperdüzlemlere kıyasla iyi bir şekilde genellenmesi beklenen yalnızca bir optimal ayırma hiperdüzlem vardır (Foody ve Mathur, 2004).

Şekil 3.1’de verileri sınıflandırmak için birçok hiperdüzlemin mevcut olduğu görülmektedir. Ancak en uygun hiperdüzlemin hangisi olduğu SVM algoritması ile belirlenebilir.



Şekil 3.1 Verileri sınıflandırmak için birçok hiperdüzlem mevcut (Jakkula, 2006)

SVM verilerin doğrusal olarak ayrılabilirliği ve doğrusal olmadığı durumlar olmak üzere iki kısımda incelenir. Şekil 3.1’de verilerin doğrusal olarak ayrıldığı durumlarda iki sınıf arasında sonsuz tane ayırıcı doğru yani hiperdüzlemler çizilebilir. Ancak bu doğrulardan sadece bir tanesi maksimum ayrılmayı sağlamaktadır. Burada SVM bu iki sınıf arasındaki mesafeyi(marjini) maksimumum yapacak doğruyu bulmaya çalışır (Ayhan ve Erdoğan, 2014). Şekil 3.2’de iki sınıfa ait verileri doğrusal olarak ayıran en uygun hiperdüzlem gösterilmiştir (Jakkula, 2006).



Şekil 3.2 Maksimum ayrılmayı sağlayan hiperdüzlem (Jakkula, 2006)

SVM verileri doğrusal olarak ayıramadığı durumlarda ise orijinal veri setini daha üst boyutta uzay düzlemine taşır. Doğrusal olmayan haritalama kullanarak veriyi taşıdığı boyutta, iki sınıf arasındaki mesafeyi maksimum yapacak ayırıcı düzlemi bulmaya çalışır (Christianini ve Taylor, 2000; Aydın ve Atasoy, 2018). Burada marjın maksimum hale getirilmesi ve yanlış sınıflandırma hatalarının ise minimum olması, C ile ifade edilen bir düzenleme parametresinin tanımlanmasıyla kontrol edilir. C parametresinin formülü “Eş. 4.2” de ifade edilmiştir (Cortes ve Vapnik, 1995; Naganna ve Deka, 2014).

3.2.5 Çok Sınıflı Destek Vektör Makineleri

SVM’ler orijinal olarak ikili sınıflandırmalar için geliştirilmiş olsalar da çok sınıflı problemleri ele alan yaklaşımları mevcuttur (Hsu ve Lin, 2002; Y. Liu ve Zheng, 2005; Madzarov vd., 2008). Çok sınıflı SVM problemi, genellikle standart SVM’nin kullanılabileceği birkaç iki sınıflı probleme ayrıştırılarak çözülür (Franc ve Hlaváč, 2002). İkinci bir yöntem ise eğitim örnekleriyle doğrudan çok sınıflı bir sınıflandırma fonksiyonu çözmektir (Weston ve Watkins, 1998). SVM’leri iki sınıflı sınıflandırıcılar olarak kullanarak çok sınıflı problemin birkaç iki sınıflı probleme dönüştürülmesi için çeşitli teknikler önerilmiştir. Bu tekniklerden bazıları;

3.2.5.1 Bire-Karşı-Bir

Algoritma N sınıflarının bütün ikili kombinasyonlarından $N(N - 1)/2$ iki sınıflı sınıflandırıcılar oluşturmaktadır. Ve her bir sınıflandırıcı örneklerin eğitilmesiyle elde edilir. Algoritma N’nin tüm ikili kombinasyonlarını kullanarak sınıflandırma yapar. Her sınıflandırıcı, örneklerle kullanılarak eğitilir (Hsu ve Lin, 2002; Madzarov vd., 2008). Tüm sınıflandırıcılar eğitildikten sonra test için bir oylama stratejisi uygulanır. Etiketlenmemiş örnek en çok oyu alan sınıfa atanır (Qian vd., 2010).

3.2.5.2 Bire-Karşı-Hepsi

Bire-karşı-hepsi yöntemi, SVM'in çok sınıflı sınıflandırıcısı için en eski uygulamasıdır. Her bir SVM'i eğitirken tüm örnekler dikkate alınır (Qian vd., 2010).

N sınıflı problemlere uygulanan bu yöntem ($N > 2$) N adet ve iki sınıflı SVM sınıflandırıcısından oluşturulur. i. SVM, i. sınıftaki örnekleri pozitif örnekler ve geri kalan tüm örnekler negatif etiketlerken eğitilir (Madzarov vd., 2008; Qian vd., 2010; Hsu ve Lin, 2002).

3.2.5.3 DAG-SVM

Yönlendirilmiş döngüsel olmayan grafik SVM (DAG-SVM Bire-Karşı-Bir yöntemiyle aynı şekilde $N(N-1)/2$ adet sınıflandırıcıyı eğitir ve her bir sınıf için model kurar (Platt vd, 2000; Qian vd., 2010).

3.2.6 Doğrusal Destek Vektör Makineleri

Doğrusal SVM problemlerinin karmaşıklığı kullanılan özelliklerin sayısına bağlı olarak değişir. SVM'in karmaşıklık düzeyi ne olursa olsun yani boyutsallık ve sınıflandırma problemleri kullanılan hiperdüzlemin düz olması ve eğri olmaması anlamında genellikle doğrusaldır (Pisner ve Schnyer, 2019).

3.2.6.1 Doğrusal Ayrılabilen Verilerde İkili Sınıflandırma

x_i girdi örneklerinin D özniteliklerine sahip olduğu, $y_i = -1$ veya $+1$ iki sınıftan birine ait olduğu bir L eğitim verisi olduğunu varsayalım. Bu durumda eğitim verileri şu formda ifade edilebilir:

$$\{x_i, y_i\} \quad \forall i \text{ için, } i = 1 \dots L, y_i \in \{-1, +1\}, x_i \in \mathbb{R}^D \quad (3.6)$$

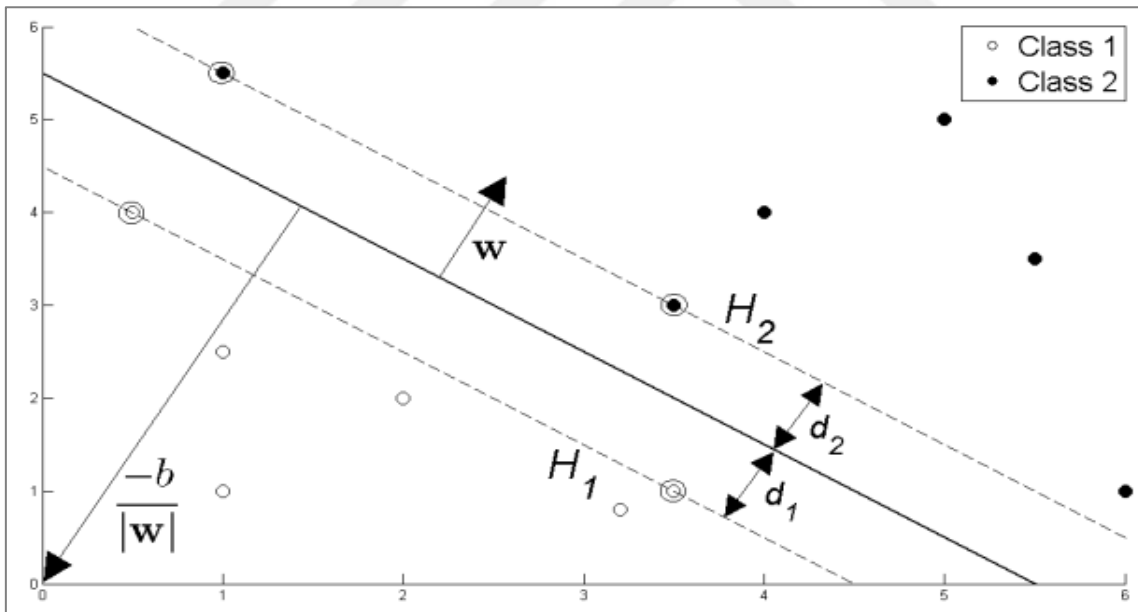
$\mathbb{R}^D$ verilerin düzlemdeki özniteliklerin sayısını ifade etmektedir. Burada eğitim verilerinin doğrusal olarak ayrılabilirdiği varsayılmıştır. Verilerin doğrusal olarak

ayrılabilirdi durumda $D=2$ olduđuunda x_1 ve x_2 grafiğine bir çizgi çizilebileceğı ve $D > 2$ olduđuunda x_1 , x_2 ve x_3 grafiğine bir hiperdüzlem oluşturabileceğı anlamına gelmektedir. Bu ayırıcı hiperdüzlem “Eş. 1.6” da olduğı gibi tanımlanabilir.

$$w \cdot x + b = 0 \quad (3.7)$$

w (weight) hiperdüzlemin normalini ifade etmektedir. $\frac{b}{\|w\|}$ ise orjinden hiperdüzleme olan dik mesafedir. Burada b (bias) sabit değıerdir (Burges, 1998a; Flake ve Lawrence, 2002; Min ve Lee, 2005; Fletcher, 2008).

Destek vektörleri ayırıcı alt düzleme en yakın olan örneklerdir. “Eş. 3.18” de belirtildiğı gibi eğitim örnekleri arasında sıfırdan büyük değıerler alan α_i lagrange değıerleri Destek vektörleri olarak geçmektedir. Burada SVM’nin görevi bu ayırıcı alt düzlemi her iki sınıfın en yakın örneklerine olabildiğince uzağıya götürmektir. Şekil 3.3’ te doğrusal ayrılabilen verilerde ayırıcı alt düzlem gösterilmiştir.



Şekil 3.3 Doğrusal olarak ayrılabilen iki sınıfa ait alt düzlem (Fletcher, 2008)

Eğitim verileri şu şekilde tanımlanabilir:

$$y_i = +1 \text{ için; } x_i \cdot w + b \geq +1 \quad (3.8)$$

$$y_i = -1 \text{ için; } x_i \cdot w + b \leq -1 \quad (3.9)$$

Bu iki denklem tek bir denklemde birleştirebilir. Bu durumda eğitim verileri için “Eş. 3.10” yazılabilir.

$$y_i(x_i \cdot w + b) - 1 \geq 0 \quad \forall_i \quad (3.10)$$

H_1 ve H_2 ayırıcı alt düzlemi ifade etmektedir. Bu durumda H_1 ve H_2 düzlemleri şu şekilde tanımlanabilir.

$$H_1 : x_i \cdot w + b = +1 \quad (3.11)$$

$$H_2 : x_i \cdot w + b = -1 \quad (3.12)$$

Şekil 3.3’te d_1 H_1 ’den orjine d_2 ise H_2 ‘den orjine olan uzaklığı ifade etmektedir. H_1 ve H_2 orjine eşit uzaklıkta yer almaktadır. Bu durumda $d_1 = d_2$ olduğu söylenebilir. Burada d_1 ve d_2 SVM’nin marjı olarak bilinmektedir. Hiper düzlemi destek vektörlerinden olabildiğince uzağa yönlendirmek için marjın maksimize edilmesi gerekmektedir. Marj $\frac{1}{\|w\|}$ ile ifade edilir. Marjın maksimize edilmesi için $\|w\|$ minimize edilmesi sağlanır. Bu durumda “Eş. 3.10”daki denklemin maksimize edilmesi için $\min\|w\|$ değerinin elde edilmesi gerekmektedir.

$\min\|w\|$ ile $\min \frac{1}{2} \|w\|^2$ eşdeğerdir. W değerini ve b değerini bulmak için karesel programlama ile “Eş. 2.2” deki optimizasyon problemi çözülmeli.

$$\min \frac{1}{2} \|w\|^2, y_i(x_i \cdot w + b) - 1 \geq 0 \quad \forall_i \quad (3.13)$$

(Fletcher, 2008).

“Eş. 3.13” deki optimizasyon probleminin çözülmesi ile destek vektörleri arasındaki uzaklık maksimize edilmiş olur, bu durumda optimal ayırıcı hiperdüzlem elde edilir. Optimal ayırıcı hiperdüzlemin bulunması için karesel optimizasyon probleminin çözümünde Lagrange çarpanlarından faydalanılmaktadır. Lagrange çarpanları bir problemin en küçüklenmesi için bu problemi dual (ikili) probleme dönüştürerek daha

hızlı çözülmesini sağlamaktadır. “Eş. 3.13” de lagrange çarpanları yardımıyla hiperdüzlem elde edilir (Ayhan ve Erdoğan, 2014).

“Eş. 3.13” ye Lagrange çarpanı α eklenecektir. $\alpha_i \geq 0 \forall_i$:

$$L_p \equiv \frac{1}{2} \|w\|^2 - \alpha[y_i(x_i \cdot w + b) - 1 \forall_i] \quad (3.14)$$

$$\equiv \frac{1}{2} \|w\|^2 - \sum_i^L a_i [y_i(x_i \cdot w + b) - 1] \quad (3.15)$$

$$\equiv \frac{1}{2} \|w\|^2 - \sum_{i=1}^L a_i y_i (x_i \cdot w + b) + \sum_{i=1}^L a_i \quad (3.16)$$

“Eş. 3.16”te w ve b ’nin minimum değerlerini, α_i ($\alpha_i \geq 0 \forall_i$)’nin ise maksimum değeri bulunmak istenmektedir. Bunun için L_p ’ye göre bu değerlerin kısmi türevleri alınacaktır.

$$\frac{\partial L_p}{\partial w} = 0 \rightarrow w = \sum_{i=1}^L a_i y_i x_i \quad (3.17)$$

$$\frac{\partial L_p}{\partial b} = 0 \rightarrow b = \sum_{i=1}^L a_i y_i = 0 \quad (3.18)$$

a_i değerinin maksimize edilmesi için “Eş. 2.6” ve “Eş. 2.7” değerleri, “Eş. 2.5”te yerine koymamız gerekmektedir. Bu durumda:

$$L_D \equiv \sum_{i=1}^L a_i - \frac{1}{2} \sum_{i,j} a_i a_j y_i y_j x_i x_j, \alpha_i \geq 0 \forall_i, \sum_{i=1}^L a_i y_i = 0 \quad (3.19)$$

$$\equiv \sum_{i=1}^L a_i - \frac{1}{2} \sum_{i,j} a_i H^+_{ij} a_j, H^+_{ij} \equiv y_i y_j x_i x_j \text{ ‘dir.} \quad (3.20)$$

$$\equiv \sum_{i=1}^L a_i - \frac{1}{2} a^T H^+ a, \alpha_i \geq 0 \forall_i, \sum_{i=1}^L a_i y_i = 0 \quad (3.21)$$

“Eş. 3.14”, “Eş. 3.16”’nin ikili bir formunu ifade etmektedir. Burada amaç L_D ’nin maksimize ve L_p ’nin minimize edilmesidir. L_D x_i girdi değerlerinin iç çarpımlarını ifade etmektedir. Optimal hiperdüzlemin elde edilmesi için “Eş. 3.16”da Dual lagrange $L_D(\alpha)$ ’yi maksimum yapacak α_i değerleri elde edilir. Dolayısıyla optimal hiperdüzlem, sıfırdan büyük değer alan Lagrange çarpanları sayesinde elde edilir. Minimum L_p ve maksimum L_D ’nin elde edilmesi için maksimum a_i değeri “Eş. 3.1” deki gibi formüle edilmektedir.

$$\text{Max}_{\alpha} \left[\sum_{i=1}^L a_i - \frac{1}{2} a^T H^+ a \right], \alpha_i \geq 0 \forall_i \text{ ve } \sum_{i=1}^L a_i y_i = 0 \quad (3.22)$$

Optimal hiperdüzlemin elde edilmesi için “Eş. 3.22”de Dual lagrange $L_D(\alpha)$ ’yi maksimum yapacak α_i değerleri elde edilir. Dolayısıyla optimal hiperdüzlem, sıfırdan büyük değer alan Lagrange çarpanları sayesinde elde edilir (Ayhan ve Erdoğan, 2014).

“Eş. 3.22” de α_i değerleri ve “Eş. 3.18” da w değerleri çözüldükten sonra geriye b değerinin hesaplanması gerekmektedir.

Destek vektörü olan x_s “Eş. 2.7”yi karşılayan herhangi bir veri noktası “Eş. 3.23” teki gibi olacaktır.

$$y_s(x_s, w + b) = 1 \quad (3.23)$$

“Eş. 3.1” de elde edilen w değeri “Eş. “3.2”de yerine koyulduğunda “Eş 3.33” elde edilir.

$$y_s(\sum_{i \in S} a_i y_i x_i \cdot x_s + b) = 1 \quad (3.23)$$

S destek vektörlerinin indislerini gösterir. S , $\alpha_i > 0$ olan i indislerini bularak belirlenir. Ardından y_s ile çarpılır. $y_s^2 = 1$ olarak alınmıştır. O halde “Eş 3.23”’in her iki tarafı y_s ile çarpılırsa “Eş. 3.24” elde edilecektir.

$$y_s^2(\sum_{m \in S} a_m y_m x_m \cdot x_s + b) = y_s \quad (3.24)$$

$$b = y_s - \sum_{m \in S} a_m y_m x_m \cdot x_s \quad (3.25)$$

Burada rastgele bir destek vektör x_s almak yerine, bütün destek vektörlerinin ortalamasının alınması daha iyi olduğu söylenebilir. Bu durumda “Eş. 3.6” elde edilir.

$$b = \frac{1}{N_s} \sum_{s \in S} (y_s - \sum_{m \in S} a_m y_m x_m \cdot x_s) \quad (3.26)$$

Optimal hiperdüzlemi elde etmemizi sağlayan w , b değişkenleri elde edilmiştir. Bundan sonra veri setine dahil olan her yeni x' noktası “Eş. 3.27” deki formüle tabi tutularak sınıflandırılır.

$$y' = \text{sgn}(w \cdot x' + b) \quad (3.27)$$

(Fletcher, 2008; Li ve ark., 2010).

3.2.6.2 Tamamen Doğrusal (Soft Marjin) Olarak Ayrılamayan Veriler İçin İkili Sınıflandırma

SVM’de veri setleri her zaman doğrusal olarak ayrılmayabilir böyle durumlarda, bazı verilerin yanlış sınıflandırılmasına müsaade edilir. Bu yanlış sınıflandırmaya izin vermek için “Eş. 3.8” ve “Eş. 3.9” ‘e ξ_i gevşek değişkeni (soft margin) eklenir. $i=1 \dots L$: “Eş. 3.7” ve “Eş. 3.8” elde edilir.

$$y_i = +1 \text{ için; } x_i \cdot w + b \geq 1 - \xi_i \quad (3.28)$$

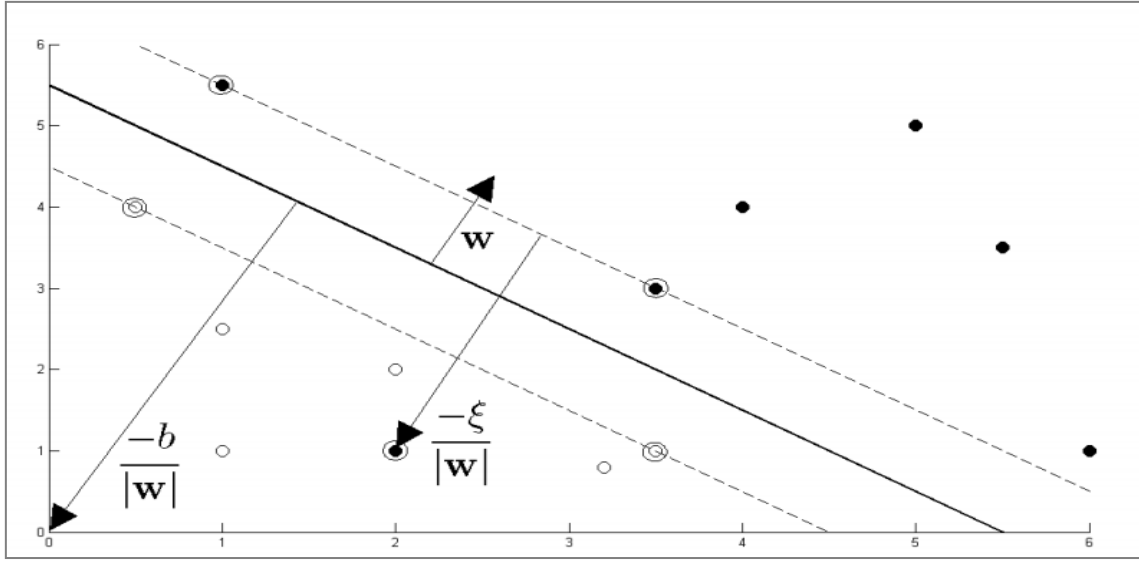
$$y_i = -1 \text{ için; } x_i \cdot w + b \leq -1 + \xi_i \quad (3.29)$$

$$\xi_i \geq 0 \quad \forall_i \quad (3.30)$$

“Eş. 3.28” ve “Eş. 3.29”u tek bir formülde birleştirdiğimizde “Eş. 4.0” elde edilir.

$$y_i(x_i \cdot w + b) - 1 + \xi_i \geq 0, \quad \xi_i \geq 0 \quad \forall_i \quad (3.31)$$

(Fletcher, 2008; Li ve ark., 2010).



Şekil 3.4 Doğrusal olarak ayrılamayan iki sınıfa ait alt düzlem (Fletcher, 2008)

Veri setinin belirli bir hata ile doğrusal ayrılmasının nedeni, veri setinin gürültülü veri içermesi, verinin çok boyutlu olması veya verinin karmaşık yapısından kaynaklanmaktadır (Ayhan ve Erdoğan, 2014). Veriler toplanırken genelde hata içerir, bu toplu hatalara gürültü adı verilir (Oğuzlar, 2004). Yumuşak marjlı SVM’de marjın yanlış tarafında yer alan veri noktalarının mesafeye göre marjdan uzaklaştıkça artan bir cezası vardır. Burada amaç yanlış sınıflandırmaların sayısını azaltmaktır. O halde “Eş 3.14” de yer alan amaç fonksiyonunu yeniden uyarlanacaktır. Bunun için “Eş. 3.32” de yer alan C ve ξ_i değerlerin bulunması gerekmektedir.

$$\min \frac{1}{2} \|w\|^2 + C \sum_{i=1}^L \xi_i \quad \text{s.t.} \quad y_i(x_i \cdot w + b) - 1 + \xi_i \geq 0 \quad \forall_i \quad (3.32)$$

Burada C kullanıcı tarafından seçilen sınıflandırma doğruluğu ve modelin karmaşıklığını kontrol eden bir parametredir. Başka bir ifadeyle yanlış sınıflandırma hatalarını minimum olmasını sağlayan bir düzenleme parametresidir. C değeri 0 ile sonsuz ($0 < C < \infty$) arasında herhangi bir değer alabilir (Kavzoğlu ve Çölkesen, 2010). C parametresi 0’a yaklaşırsa gevşek değerler kısıtsız hale gelir. Bu da marjın genişliğinin maksimum hale gelmesi anlamına gelir. Böyle bir durumda marjın içinde veya yanlış tarafta kaç tane veri olduğu önemsiz bir duruma gelir. Bu durumun oluşturduğu avantaj; yeni veri kabulünde başarı artarken, dezavantajı ise; modelin öğrenme fazında

hassasiyetin azalıyor olmasıdır (Fletcher, 2008). C'nin büyük olması hatalara daha büyük bir ceza verilmesi anlamına gelir. Yani C parametresi sonsuza yaklaştıkça modelin hassasiyeti artar bu da marjın genişliğini azaltır. Bu durum öğrenme fazında hassasiyeti yükseltir, fakat yeni veri kabulünde zorluklar yaşanır. ξ_i eğitim verilerindeki hata sayılarına bağlıdır. Bu nedenle hatalar için ekstra bir maliyet belirlemek için “Eş. 3.32”de yer alan amaç fonksiyonunun minimize edilmesi gerekmektedir (Burges, 1998; Fletcher, 2008; Liu vd., 2013; Güran vd., 2014). Amaç fonksiyonunun Lagrange çarpanı olan α_i 'yi kullanarak maksimize, w, b ve ξ_i 'ye göre minimize edilir. $\alpha_i \geq 0, \mu_i \geq 0 \forall_i$:

$$L_p \equiv \frac{1}{2} \|w\|^2 + C \sum_{i=1}^L \xi_i - \sum_i^L a_i [y_i(x_i \cdot w + b) - 1 + \xi_i] - \sum_{i=1}^L \mu_i \xi_i \quad (3.33)$$

“Eş. 4.2” de w, b ve ξ_i 'ye göre kısmi türevleri alınarak sıfıra eşitlenir.

$$\frac{\partial L_p}{\partial w} = 0 \rightarrow w = \sum_{i=1}^L a_i y_i x_i \quad (3.34)$$

$$\frac{\partial L_p}{\partial b} = 0 \rightarrow \sum_{i=1}^L a_i y_i = 0 \quad (3.35)$$

$$\frac{\partial L_p}{\partial \xi_i} = 0 \rightarrow C = a_i + \mu_i \quad (3.36)$$

L_D daha öncede belirtilen “Eş. 3.0” ile aynı formdadır. “Eşitlik 3.37” da yer alan $\mu_i \geq 0 \forall_i$ olacaktır. Buradan da $a_i \geq C$ anlamına gelir. “Eş. 4.6” a_i değerinin maksimize eder.

$$\text{Max}_{\alpha} \left[\sum_{i=1}^L a_i - \frac{1}{2} a^T H a \right] \text{ s.t. } 0 \leq a_i \leq C \forall_i \text{ ve } \sum_{i=1}^L a_i y_i = 0 \quad (3.37)$$

(Fletcher, 2008).

3.2.7 Doğrusal Olmayan Destek Vektör Makineleri

Doğrusal olmayan SVM verilerin doğrusal olarak ayrılmadığı durumlarda, çekirdek fonksiyonları yardımıyla bu verileri daha üst boyutlu uzaya taşıyarak verilerin doğrusal bir şekilde ayrılmasına olanak tanımaktadır (Demirel vd., 2020). Gerçek hayat problemlerinde bir veri setinin ayırıcı alt düzlem ile doğrusal ayrılması her zaman mümkün değildir (Ayhan ve Erdoğan, 2014). SVM’yi doğrusal ayrılabilen verilere uygularken “Eş. 4.7” da girdi değişkenlerinin iç çarpımlarından bir **H** matrisi oluşturulur.

$$H_{ij} = y_i y_j k(x_i x_j) = x_i x_j = x_i^T x_j \quad (3.38)$$

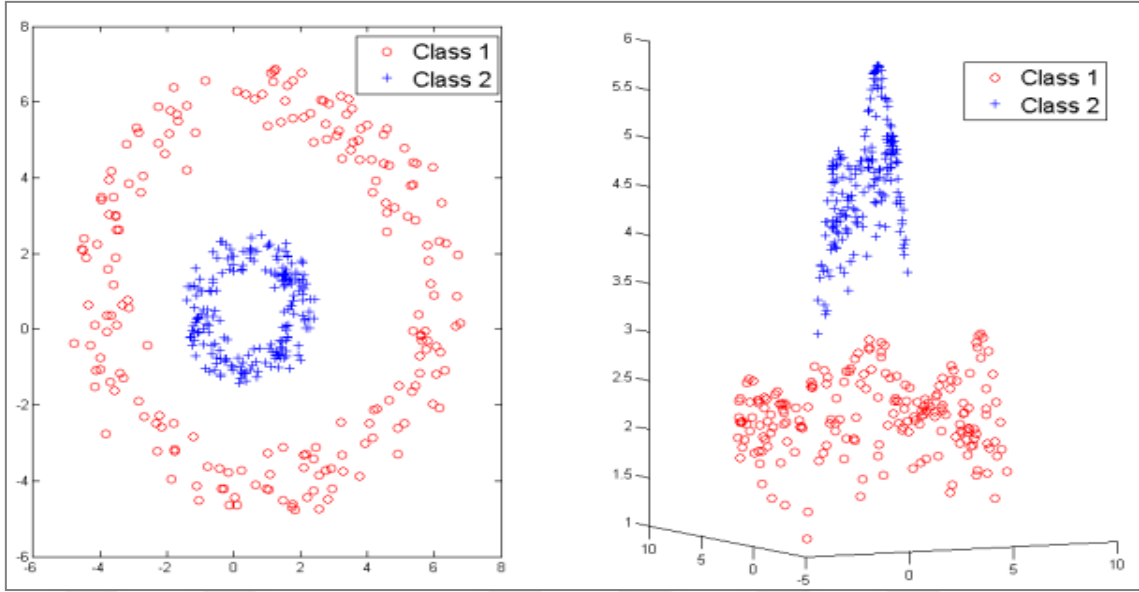
$k(x_i x_j)$ çekirdek fonksiyonlarını ifade eder. Burada $k(x_i x_j) = x_i x_j = x_i^T x_j$ doğrusal çekirdek fonksiyonudur. Şekil 4’te gösterildiği gibi doğrusal olmayan haritalama yöntemi, x_i girdi verilerini daha üst boyuta taşıyarak, burada verilerin doğrusal ayrımını gerçekleştirir. “Eş. 4.8”te 2 boyutlu girdi vektörleri (x_1, x_2) ve 3 boyutlu özellik uzayı (z_1, z_2, z_3) olmak üzere fonksiyonların özellik uzayı,

$$\varphi: R^2 \rightarrow R^3 (x_1, x_2) \rightarrow (z_1, z_2, z_3) := (x_1^2, \sqrt{2x_1 x_2}, x_2^2) \quad (3.39)$$

yukarıdaki eşitlik ile ifade edilir. Özellik uzayında haritalanmış girdi vektörlerinin iç çarpımları “Eş. 3.40”taki gibi elde edilir.

$$(x_1^2 x_1'^2 + 2x_1^2 x_1' x_2'^2 + x_2^2 x_2'^2)^2 \quad (3.40)$$

(Fletcher, 2008; Ayhan ve Erdoğan, 2014). Doğrusal ayrılamayan veriler Şekil 3’te gösterilmiştir.



Şekil 3.5 Doğrusal ayrılamayan veriler (Fletcher, 2008)

3.2.8 Çekirdek Fonksiyonları

Çekirdek fonksiyonları iki vektörün iç çarpımlarının hesaplanmasına dayanmaktadır. Çekirdek fonksiyonları doğrusal ayrılamayan x girdi vektörlerini doğrusal olmayan haritalama yöntemi kullanarak $\varphi(x)$ özellik vektörüne dönüştürür ve daha yüksek boyutta yeniden biçimlendirmeye çalışır. $\varphi(x)$ özellik haritası olarak tanımlanır “Eş. 5.0”de gösterilmiştir (Fletcher, 2008; Ayhan ve Erdoğan, 2014). Çekirdek çoğunlukla yüksek boyutlu özellik uzayında bir noktanın çarpımına denk gelir (Hofmann vd., 2008). Çekirdek, veri doğrusal ise verileri bölmek için bir ayırıcı hiperdüzlem kullanılabilir. Ancak çoğu zaman veriler doğrusal ayrılamamaktadır ve veri kümelerinin birbirinden ayrılmaz olduğu bir durum mevcuttur. Bu durumda girdi verilerini yüksek boyutlu bir uzaya doğrusal olmayan bir şekilde eşlemek için çekirdek kullanılır. Bu sayede eşleme doğrusal ayrılabilir hale gelir (Jakkula, 2006). Çekirdek fonksiyonlarında temel fikir doğrusal bir algoritma verildiğinde, önce giriş uzayında yer alan X verileri doğrusal olmayan eşleme $\varphi(\cdot)$ aracılığıyla H 'ye (Hilbert Uzayı) yani özellik uzayına eşlenir. $X \rightarrow H$ akabinde algoritma verinin vektör temsili olan $\varphi(x)$ üzerinde çalıştırır. $\Phi(\cdot)$ haritasının amacı, verilerin doğrusal olmayan yapılarını H 'deki doğrusal yapıya çevirmektir (Karatzoglou ve ark., 2004; Pérez-Cruz ve Bousquet, 2004). SVM'de çekirdek fonksiyonu ve çekirdek fonksiyonuna ait uygun parametre

değerinin seçimi SVM'in sınıflandırma performansını etkileyen önemli bir faktördür (Smola vd., 1998; Güran vd., 2014). Çekirdek fonksiyonu “Eş. 3.41”de ki formda gösterilmiştir.

$$k: X \times X \rightarrow \mathbb{R}, (x, x') \mapsto k(x, x') \text{ tüm } x, x' \in X \text{ olsun.}$$

$$k(x, x') = \langle \varphi(x), \varphi(x') \rangle \quad (3.41)$$

k benzerlik ölçüsü, genelde çekirdek olarak tanımlanır (Ben-Hur vd., 2008; Hofmann vd., 2008; Lampert, 2009; Nanda vd., 2018).

3.2.8.1 Çekirdekleştirme

Çekirdekleştirme, genelleştirilmiş maksimum marj sınıflandırıcısını bulmamızı sağlayan optimizasyon probleminin boyutunu küçültmemizi sağlar. Fakat $\phi(x_1), \dots, \phi(x_n)$ değerlerinin saklanması ve hesaplanması çok fazla bellek ihtiyacı gerektirmektedir. Bu durumdan kurtulmak için, $\phi(x_i)$ dönüştürülmüş özniteliklerin iç çarpımlarda sadece ikili formda $\phi(x_i), \phi(x_j)$ $i, j = 1, \dots, n$ şeklinde ifade ederek $\phi(x_i)$ 'nin fazla bellek sorunu çözülmüş olur.

$X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$, $Y = \{y_1, \dots, y_n\} \subset \{-1, +1\}$ etiketli eğitim verileri, $k: \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ bir çekirdek işlevi olsun. $C > 0$, çekirdekleştirilmiş maksimum marj sınıflandırıcısının karar fonksiyonu “Eş. 3.42” deki gibi ifade edilir.

$$f(x) = \sum_{i=1}^n a_i^* k(x_i, x) \quad (3.42)$$

$a_1^*, \dots, a_n^*$ katsayıları çözülerek elde edilen karar fonksiyonu aşağıdaki gibidir;

$$\min_{\substack{a_1^*, \dots, a_n^* \in \mathbb{R} \\ \xi_1, \dots, \xi_n \in \mathbb{R}^+}} \sum_{i,j=1}^n a_i^* a_j^* k(x_i, x_j) + \frac{C}{n} \sum_{i=1}^n \xi_i \quad (3.43)$$

(Gärtner, 2003; Lampert, 2009).

Çekirdek yöntemlerinin iki bileşeni mevcuttur. Bunlar çekirdek makinesi ve çekirdek işlevidir. Çekirdek makinesi, öğrenme görevini ve çözümün aranma yöntemini

kapsar. Çekirdek işlevi ise hipotezi yani olası çözümlerin nasıl oluşturulduğunu göstermektedir (Gärtner, 2003).

Literatürde sınıflandırma problemi için sıkça kullanılan çekirdek fonksiyonları; Doğrusal Çekirdek, Radial Basis Çekirdek, Polinom Çekirdek, ANOVA, Sigmoid Pearson VII işlevsel çekirdek fonksiyonlarıdır. Söz konusu fonksiyonlar;

3.2.8.2 Doğrusal Çekirdek Fonksiyonu

Genellikle metin kategorizasyonunda sıkça tercih edilmektedir. Veriler hakkında önceden bilgi sahibi olmadığımız durumlarda kullanılır. $x_i, x_j \in \mathbb{R}^n$

$$k(x_i, x_j) = x_i x_j \quad (3.44)$$

(Fletcher, 2008; Kasnavi vd., 2018; Nanda vd., 2018).

3.2.8.3 Radial Basis Çekirdek Fonksiyonu (RBF)

$$k(x_i, x_j) = e^{-\left(\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)} \quad (3.45)$$

RBF sınıflandırmada en çok tercih edilen çekirdek fonksiyonları arasında yer almaktadır. SVM sınıflandırma veya tahmin yaparken RBF çekirdeği sayesinde rastgele veri noktalarının, veri kümesini parçalayabilir (Cho vd., 2008; Fletcher, 2008). Eğer σ çok küçükse, uzaklıkları σ 'ya yakın olan örnekler etkilenecektir. Aksi durumda, uzaklıkları σ 'dan çok daha büyük olan örnekler mevcutsa ifadenin değeri kademeli olarak sıfıra yaklaştırılır (Gartner, 2008; Hofmann vd., 2008; Song vd., 2008).

3.2.8.4 Polinom Çekirdek Fonksiyonu

Görüntü işlemede genellikle kullanılan bir çekirdek fonksiyonudur.

$$k(x_i, x_j) = (x_i \cdot x_j + a^*)^{d^*} \quad (3.46)$$

d^* ve a^* polinom çekirdeğinin parametresidir.

3.2.8.5 Sigmoid Çekirdek Fonksiyonu

SVM için popüler çekirdek fonksiyonu olan sigmoid çekirdek, temeli sinir ağlarından gelmektedir. “Eş 3.47” ile ifade edilmektedir.

$$k(x_i, x_j) = \tanh(a^* x_i \cdot x_j - b^*) \quad (3.47)$$

a^*, b^* sigmoid çekirdeğinin parametrelerini ifade eder (Lin ve Lin, 2003).

3.2.8.6 ANOVA Çekirdek Fonksiyonu

ANOVA çekirdeklerini oluşturmak için bazı G kümeleri $X = G^N$ ve $G \times G$ üzerindeki k çekirdekleri dikkate alınır. Burada $i = 1, \dots, N$ için $P = 1, \dots, N$ ANOVA çekirdeği “Eş. 5.7”deki gibi oluşturulur.

$$k_P(x, x') := \sum_{1 \leq i_1 < \dots < i_P \leq N} \prod_{p=1}^P k^{i_p}(x_{i_p}, x'_{i_p}) \quad (3.48)$$

(Hofmann ve ark., 2008).

3.2.8.7 Pearson VII Evrensel İşlev Çekirdeği

Pearson VII fonksiyonunun genel formu “Eş. 3.49” deki gibi ifade edilmektedir.

$$f(x) = \frac{H^{**}}{\left[1 + ((2(x - x_0)\sqrt{2^{1/\omega} - 1})/\delta)^2\right]^\omega} \quad (3.49)$$

Bu eşitlikte H^{**} tepenin x_0 merkezindeki tepe yüksekliğidir. x bağımsız değişkeni ifade etmektedir. δ parametredir. ω ise tepe noktasının yarı genişliğini ifade etmektedir.

$$k(x_i, x_j) = \frac{1}{\left[1 + ((2|x_i - x_j|\sqrt{2^{1/\omega} - 1})/\delta)^2\right]^\omega} \quad (3.50)$$

x_i, x_j iki vektörün bağımsız değişkenleridir. Denklemdaki tepe noktasın kaldırılır ve H^{**} 'in 1 ile değiştirilir buda çekirdeğin genelleştirmesinde bir değişiklik oluşturmamaktadır. Bu şekilde pearson VII işlev çekirdeği köşegen üzerinde olanlar ve herhangi bir rastgele x_i, x_j çifti için 0 ile 1 değerleri arasında değişen tüm girişlerde simetrik bir matrise yol açmaktadır. Pearson VII işlev çekirdeği hem sağlam hem de SVM'lerin eşit veya daha iyi genelleştirme performansına yol açan standart çekirdek işlevleriyle karşılaştırıldığında eşit veya daha güçlü eşleme gücüne sahiptir (Zhang ve Ge, 2013).

3.2.8.8 String Çekirdek Fonksiyonu

İki çeşit string çekirdek mevcuttur. Bunlardan ilki uzun dizgenin tüm sürekli alt dizinlerini öznitelikler olarak çıkarmaktadır. İkincisi ise; dizi çekirdeği adı verilen sürekli olmayan tüm alt dizileri veya alt dizileri kullanmaktadır. String çekirdeklerde daha düşük işlem karmaşıklığına ve daha fazla bilgi işlem verimliliğine sahip oldukları ilk tür dizi çekirdeği ele alınmıştır. String çekirdek için bazı gösterimler ve tanımlar oluşturmamız gerekmektedir (Kaynak).

Σ bir dizi karakter içeren sonlu bir alfabe olsun. Σ^* Σ üzerindeki tüm boş olmayan dizileri kümesidir.

$|x|$ gösterimi bir string olan x 'in boyutunu göstermektedir. x dizisinin uzunluğunu göstermek için u 'dan faydalanılır. $u \sqsubseteq x$. u, x 'in bir alt dizinin göstermektedir. incelenecek dizi çekirdeği “Eş. 3.51” da gösterilmiştir.

$$K(x, x') = \sum_{s \in \Sigma, s' \in x'} w_s \delta_{s, s'} = \sum_{s \in \Sigma^*} w_s num_s(x) num_s(x') \quad (3.51)$$

$num_s(x)$ x 'in bir alt dizi olarak s sayısını belirtmektedir.

İlgili parametrelerin doğrusal ayrılamayan verileri doğrusal olarak ayrılabilen özellik alanına eşleştirmesi gerekmektedir (Karatzoglou ve ark., 2006; Fletcher, 2008).

Bu durumda ilk önce çekirdek fonksiyonun seçilmesi gerekmektedir. Daha sonra x girdi verileri $\varphi(x)$ özellik alanı ile eşleştirilmelidir. $(x \rightarrow \varphi(x))$ sınıflandırma yapılması için H matrisi “Eş. 3.52”deki gibi oluşturulur.

$$H = y_i y_j \varphi(x_i) \cdot \varphi(x_j) \quad (3.52)$$

C değerini seçtikten sonra, Lagrange çarpanı olan α değeri “Eş. 5.2”de ki gibi elde edilecektir.

$$\sum_{i=1}^L a_i - \frac{1}{2} a^T H \alpha \quad (3.53)$$

Maksimize edilecek α çarpanının “Eş. 3.54”e tabi tutulması gerekmektedir.

$$0 \leq \alpha_i \leq C \quad \forall_i \text{ ve } \sum_{i=1}^L \alpha_i y_i = 0 \quad (3.54)$$

Bu işlemler karesel programlama yöntemine tabi tutularak yapılır. w değeri “Eş. 3.55” teki gibi elde edilir.

$$w = \sum_{i=1}^L \alpha_i y_i \varphi(x_i) \quad (3.55)$$

Destek Vektörleri olan S kümesini belirlemek için $0 \leq \alpha_i \leq C$ indisler bulunmalıdır. Geriye b değerinin hesaplanması kaldı. “Eş. 3.56”te b değerini elde edilmiştir.

$$b = \frac{1}{N_s} \sum_{s \in S} (y_s - \sum_{m \in S} a_m y_m \varphi(x_m) \cdot \varphi(x_s)) \quad (3.56)$$

Her yeni x' noktası $y' = \text{sgn}(w \cdot \varphi(x') + b)$ modeline tabi tutularak sınıflandırılır (Fletcher, 2008).

SVM’ de çekirdek fonksiyonu ve çekirdek fonksiyonuna ait uygun parametre değerinin seçimi, SVM’ nin sınıflandırma performansını etkileyen önemli bir faktördür. Belirtilen faktörler probleme uygun seçilmediği takdirde SVM’nin genelleştirme performansı olumsuz yönde etkilenecektir (Güran vd., 2014).

3.2.9 PCA Analizi ile Destek Vektör Makinesi Formülasyonu ve Çekirdek Fonksiyonu

Temel Bileşen Analizi (Principal Component Analysis - PCA) çok değişkenli istatistiksel bir tekniktir (Abdi ve Williams, 2010). Çok değişkenli veri analizinin temelini oluşturmaktadır (Wold vd., 1987). Tüm bilimsel disiplinler tarafından kullanılmaktadır. En eski çok değişkenli teknikler arasında yer almaktadır. Gözlemlerin birbiriyle ilişkili birkaç nicel bağımlı değişken tarafında tanımlandığı bir veri tablosunu analiz eder. Amacı, tablodan önemli bilgileri çıkarmak, onu temel bileşenler adı verilen bir dizi yeni ortogonal değişken olarak temsil etmek ve gözlemlerin ve değişkenlerin benzerlik modelini haritalarda noktalar olarak göstermektir (Abdi ve Williams, 2010). PCA'nın kökenleri çok değişkenli veri analizinden gelmektedir. En yaygın kullanım alanlarından biri, büyük veri setlerini analiz etmeye çalışmanın ilk adımı olmasıdır. Diğer yaygın uygulamalardan bazıları, gürültü giderme sinyalleri ve veri sıkıştırmasıdır. Temel olarak PCA, büyük veri kümelerinin boyutsallığını azaltmak için bir vektör uzayı dönüşümü kullanır (Richardson, 2009).

İkili değişkenlerde PCA basit ve anlaşılır bir primal-dual destek vektör makinesi formülasyonu sunmaktadır. Yüksek boyutlu bir özellik uzayına eşleme ve çekirdek hilesinin (Mercer Teoremi) uygulaması düşünülerek çekirdek PCA elde edilmiştir.

3.2.9.1 Çekirdek PCA

Özellik alanındaki nokta ürünlerini hesaplamak için çekirdek yöntemi sadece SVM'lerle sınırlı değildir. PCA gibi nokta çarpımlar cinsinden oluşturulabilen algoritmaların lineer olmayan genellemelerini geliştirmek için kullanılabileceği belirtilmiştir. PCA en yaygın özellik çıkarma algoritmaları arasında yer almaktadır. Özellik alanındaki PCA, çekirdek PCA olarak tanımlanan bir algoritmayı doğurur. Bu algoritma bir özdeğer problemini çözerek lineer olmayan özellik çıkarma fonksiyonlarını hesaplar.

$$f_n(x) = \sum_{i=1}^m q_i^n k(x_i, x) \quad (3.57)$$

Burada q_i^n normalleştirme terimidir. $K_i := (k(x_i, x))$ çekirdek matrisinin n 'inci özvektörünün bileşenlerini ifade etmektedir. Kısaca durumu özetleyecek olursak Hilbert uzayında PCA yapmak için, özvektörleri $\mathbf{v}$ ve özdeğer λ , öznilelik uzayında ise kovaryans matrisinin özdeğerlerini bulmamız gerekmektedir.

$$\text{Cov} := \frac{1}{m} \sum_{i=1}^m \Phi(x_i) \Phi(x_i)^T \quad (3.58)$$

Burada $\Phi(x_i)^T$, $\Phi(x_i)$ 'nin transpozudur. Transpozu almanın nedeni çok yüksek boyutlu durumlarda hesaplama maliyetini azaltmaktır.

$$\text{Cov} = \lambda \mathbf{v} \quad (3.59)$$

$\lambda \neq 0$ eğitim verileri Φ görüntülerinin olduğu aralıkta yer almalıdır. Bu sayede $\mathbf{v}$ çözümünü şu şekilde genişletebiliriz.

$$\mathbf{v} = \sum_{i=1}^m q_i \Phi(x_i) \quad (3.60)$$

Böylece sorun q_i 'yi bulmaya kadar indirgemiş olduk. Bu durum genişleme katsayıları için ikili bir özdeğer bulma sorununa yol açtığı görülmektedir.

$$m\lambda q = Kq \quad (3.61)$$

Burada $q = (q_1, \dots, q_m)^T$ x test noktasında lineer olmayan özellikleri çıkarmak için, özellik uzayında $\Phi(x)$ ile n 'inci normalleştirilmiş özvektör arasındaki nokta çarpımı hesaplanacaktır.

$$\langle \mathbf{v}^n, \Phi(x) \rangle = \sum_{i=1}^m q_i^n k(x_i, x) \quad (3.62)$$

Bu yöntem, nokta çarpımlarını özellik alanında açıkça almaktan hesaplama açısından çok daha ucuz olacaktır (Schölkopf ve Smola, 2002).

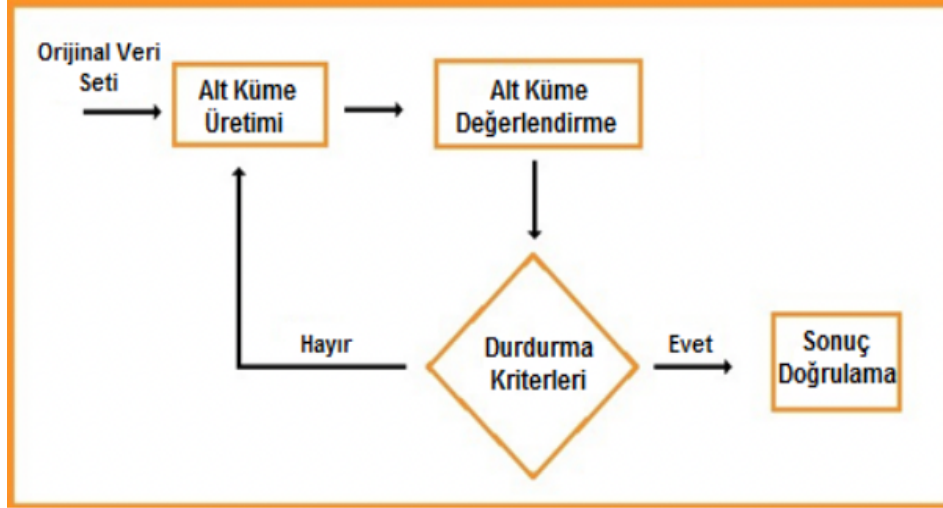
3.2.10 Özellik Seçimi (Future Selection)

Özellik Seçimi temelinde orijinal veri setini belirleyecek olan alt kümenin seçilmesi esasına dayanır. Özellik seçiminde, kullanılacak olan algoritmaya göre veri setindeki özellikleri değerlendirip n adet özellik arasında k adet özelliği seçer. Özellik seçimindeki amaç; ilgili problem için faydalı olabilecek özellikleri seçerek, veri kümesindeki özellik sayısını azaltır. Özellik sayının azaltılmasının uygulama alanında birçok avantajı mevcuttur. Bu avantajlar:

- Özellik kümesinin boyutunu azaltıp, algoritmanın daha hızlı çalışmasını sağlar.
- Alakasız veri ve gürültülü verileri ortadan kaldırır.
- Verinin daha kaliteli olmasını sağlar.
- Veri kümesinin daha basit anlaşılabilir, tanımlanabilir ve görselleştirilebilir hale gelmesi sağlar.
- Verinin depolanması için gerekli hafıza miktarını azaltır.
- Veri oluşturulurken gerekli olan verinin toplanmasında kaynak tasarrufu sağlar.
- Mevcut modelini başarının artmasını sağlar (Saeys vd., 2007; Budak, 2018).

3.2.10.1 Özellik Seçiminin Adımları

Şekil 3.6'da özellik seçiminin genel adımları gösterilmiştir. Burada orijinal veri setinden bir alt özellik kümesi oluşturulur. Akabinde ilgilenilen özellikler için farklı formüller yardımıyla değerlendirme yapılarak ilgili özelliğin seçilip seçilmeyeceğine karar verilir. Seçilen özellik ilgili alt kümeye dahil edilir ve algoritmanın durdurma kriteri sağlanana kadar bu süreç devam eder (Budak, 2018).



Şekil 3.6 Özellik seçimi akış şeması (Budak, 2018)

3.2.10.2 Alt Küme Üretimi

Arama uzayındaki aday alt kümelerin belirlenmesine yardımcı olan sezgisel bir süreçtir. Bu süreç iki temel konu ile belirlenmektedir ilk konu, arama başlangıç noktasıdır. Arama yönü ileri yönlü geri yönlü veya iki yönlü olabilir. İkinci konu, kullanılacak arama stratejisine karar verilmesidir. N adet özelliğin bulunduğu veri setinin 2^N adet aday alt kümesi bulunmaktadır. N sayının sürekli artması arama işlemini engelleyebilecek düzeye ulaşır. Bundan dolayı, komple arama, ardışık arama ve rastgele arama gibi stratejiler geliştirilmiştir (Liu, 2005; Budak, 2018).

3.2.10.3 Özellik Seçim Yöntemleri

Özellik seçiminde kullanılacak olan yöntemler üç grupta incelenecektir. Bunlar, İstatistiksel bilgiye dayanan filtreleme (filter) yöntemi, özellikler üzerinde arama işlemlerini gerçekleştiren sarmal (wrapper) yöntemi ve en iyi bölen ölçütünü bulmayı sağlayan gömülü (embedded) yöntemlerdir. Filtreleme yöntemlerinde veri madenciliği algoritma çalışmadan önce özellik seçimi yapılmaktadır. Sarmal yöntemlerde veri madenciliği algoritması, en iyi özelliklerin seçimi için araç kullanılır. Gömülü yöntemlerde ise veri madenciliği algoritması ve özellik seçimi algoritması eş zamanlı olarak çalışmaktadır. Özellik seçiminin yaygın olarak kullanıldığı alanlar; metin

madenciliği, kanser teşhisi, sahtecilik tespiti, kredi puanlama, müşteri kaybı analizi, spam e-posta analizi gibi alanlardır (Liu, 2005; Budak, 2018).

3.2.11 Performans ve Tahmin Ölçütleri

3.2.11.1 Karışıklık Matrisi

SVM algoritmasının performansının ölçülmesinde kullanılan bazı ölçütler aşağıdaki gibidir.

TP: Doğru Pozitif (True Positive): Bu ölçüt model tarafından doğru olan ve gerçekten de doğru tahmin edilen örneklerin sayısını ifade eder.

TN: (Doğru Pozitif (True Negative): Model tarafından, gerçekte doğru olmayan doğru sınıflandırılan negatif örneklerin sayısını ifade etmektedir.

FP: Yanlış Pozitif (False Positive): Doğru grupta yer alıp fakat model tarafından yanlış sınıflandırılan negatif örnekleri ifade eder.

FN: False Negatif (False Negative): Yanlış sınıfta olup model tarafından pozitif sınıflandırılan örnekleri ifade eder.

Doğruluk (Accuracy): Doğru sınıflandırılan örneklerin sayısının tüm örneklerin sayısına bölünmesiyle elde edilir. $Accuracy = \frac{TN + TP}{TN + FP + TP + FN}$ şekliden ifade edilir (Hossin ve Sulaiman, 2015).

3.2.11.2 Kappa İstatistiği

Cohen tarafından önerilen Cohen's kappa istatistiği modelin güvenilirliği ve sınıfların birbiriyle uyumlu olup olmadığını değerlendirmek için kullanılmaktadır. Cohen kappa istatistiği iki sınıflı verilerde kullanılır. -1 ile +1 arasında değer alan Cohen's kappa istatistiği, kolay hesaplanması ve pratik yorumlanması ayrıca şansa beklenen uyumu düzeltmesi bakımından önemli bir avantaj sağlamaktadır (Bilgen ve Doğa, 2017).

$$\text{Cohen's kappa: } \kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \quad (3.63)$$

$\bar{P}$ gözlenen uyumluluk oranı, $\bar{P}_e$ ise tesadüf/şansa bağlı uyumluluk oranını belirtmektedir.

Kappa istatistiğinin (κ) pozitif olması uyumun şansa beklenen uyumundan fazla olduğunu, negatif olması ise şansa uyumun beklenenden az olduğunu göstermektedir (Bilgen ve Doğa, 2017). Kappa istatistiğinin yorumlanmasına ilişkin değerler Şekil 3.7’de gösterilmiştir.

κ	Uyumun Gücü
$< 0,00$	Zayıf
$0,00 - 0,20$	Önemsiz
$0,21 - 0,40$	Düşük
$0,41 - 0,60$	Orta
$0,61 - 0,80$	Önemli
$0,81 - 1,00$	Çok Yüksek

Şekil 3.7 Kappa istatistiğinin yorumlanmasına ilişkin değer aralıkları (Bilgen ve Doğa, 2017)

3.2.11.3 ROC eğrisi

ROC eğrileri sonuç değişkenlerinin iki olasılıklı (var, yok, iyi, huylu, kötü huylu vb.) durumlarda kullanılmaktadır. ROC eğrileri (x,y) eksenlerinde (0,0) ile (1,1) değerleri arasında artan bir fonksiyondur. x eksenin de yanlış pozitif (FP yukarıda böyle ifade etmişsin), y eksenin de ise doğru pozitif değerleri yer almaktadır. ROC eğrileri ile yapılan analizler sonrasında modelin başarısının iyi mi yoksa kötü mü olduğunun kararı verilir. ROC eğrisi sol üst köşeye ((0,1) noktası) yaklaştıkça modelin ayırt etme gücü artar. Bu durumda ROC eğrisi altında kalan alan (area under the curve =AUC) ne kadar büyükse modelin sınıflama başarısının da yüksek olduğu söylenir. AUC en büyük değeri 1 en küçük değeri ise 0,5’tir (Selim, 2013; Çiğdem, 2016).



4. BULGULAR

Günümüzde modern kuruluşlar yoğun bir veri bombardımanı altındadır. Fakat bu veriler işleme tabi tutulmadığı sürece ham veriler bir değer sunmamaktadır. Makine öğrenmesi algoritmalarının verileri analiz etmesi ve bu analiz sonucunda verilerin yorumlanması önem arz etmektedir Makine öğrenmesi algoritmalarından biri olan SVM verileri sınıflandırmak ve verilerden anlamlı bir sonuç çıkarmak için çekirdek

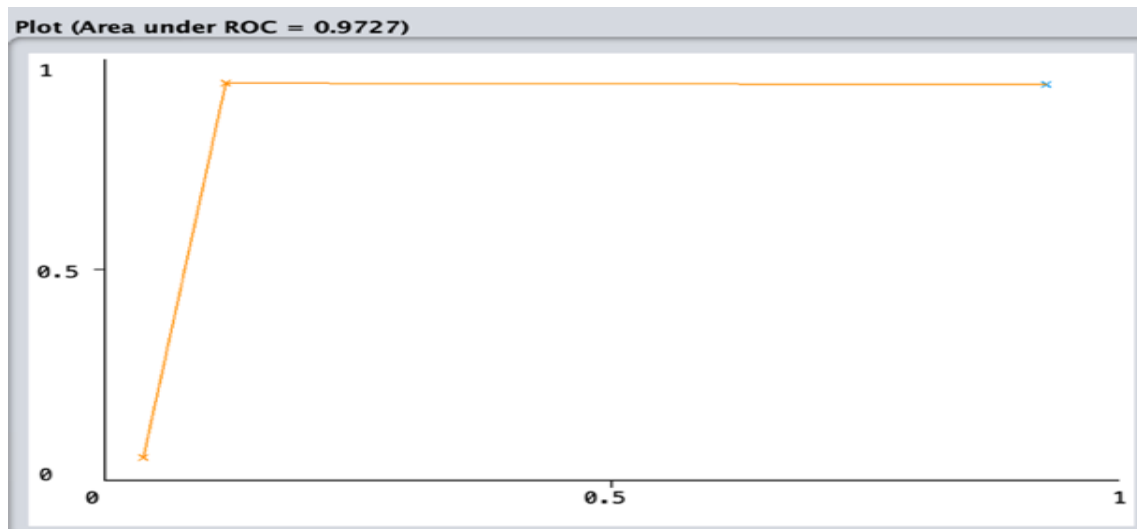
4.1 Göğüs Kanseri Veri Seti Analiz Sonuçları

Lineer çekirdek fonksiyonu için C parametresinin değeri bir (1) alınmıştır Lineer çekirdek fonksiyonu ile analiz edilen verilere ait sonuç Çizelge 4.1’de gösterilmiştir

Çizelge 4.1 Göğüs kanseri veri seti Lineer çekirdek fonksiyonu analiz sonucu

iyi huylu		kötü huylu		kappa istatistik değeri	doğruluk oranı
tp	tn	tp	tn		
356	1	201	11	0.973	97.891

Çizelge 4.1’de lineer çekirdek fonksiyonu yardımıyla analiz edilen verilerde doğruluk oranı 97.891 çıkmıştır. Ayrıca kappa istatistik değerinin 0.973 çıkması modelin güvenilir ve sınıflar arasındaki uyumun çok iyi olduğu söylenebilir.



Şekil 4.1 Göğüs kanseri veri seti lineer çekirdek fonksiyonu ile analiz edilen örneklere ait ROC eğrisi

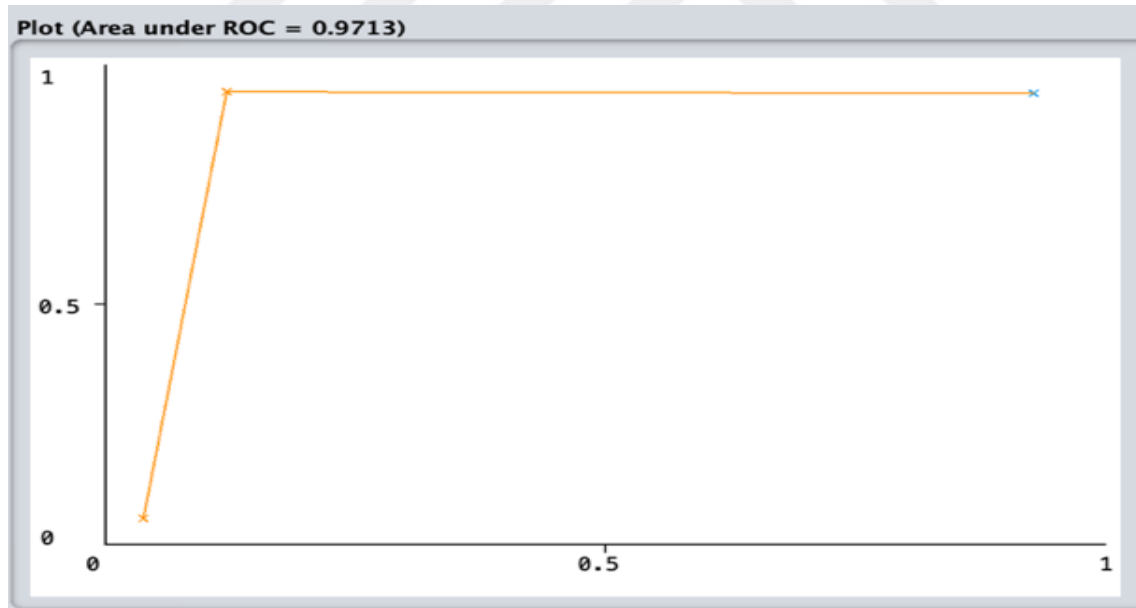
Şekil 4.1’de iyi huylu ve kötü huylu örneklere ait ROC eğrisinin altında kalan alanın büyük olduğu görülmektedir. Ayrıca ROC değeri 0.9727 elde edilmiştir. Modelin örnekleri sınıflandırma başarısının iyi olduğu söylenebilir.

Polinom çekirdek fonksiyonu için C parametresinin değeri on bin ($C=10000$) seçilmiştir.

Çizelge 4.2 Göğüs kanseri veri seti polinom çekirdek fonksiyonu analiz sonucu

iyi huylu		kötü huylu		kappa istatistik değeri	doğruluk oranı
tp	tn	tp	tn		
355	2	201	11	0.9507	97.7153

Çizelge 4.2’de polinom çekirdek fonksiyonu yardımıyla analiz edilen verilerde doğruluk 97.7153 çıkmıştır. Ayrıca kappa istatistik değerinin 0.9507 çıkması polinom çekirdek fonksiyonu ile kurulan modelin güvenilir olduğu sınıflar arasındaki uyumun yüksek olduğu söylenebilir.



Şekil 4.2 Göğüs kanseri veri seti polinom çekirdek fonksiyonu ile analiz edilen örneklere ait ROC eğrisi

Şekil 4.2’de polinom çekirdek fonksiyonuyla analiz edilen iyi huylu ve kötü örneklere ait ROC eğrisinin altında kalan alanın büyük olduğu görülmüştür. Ayrıca ROC

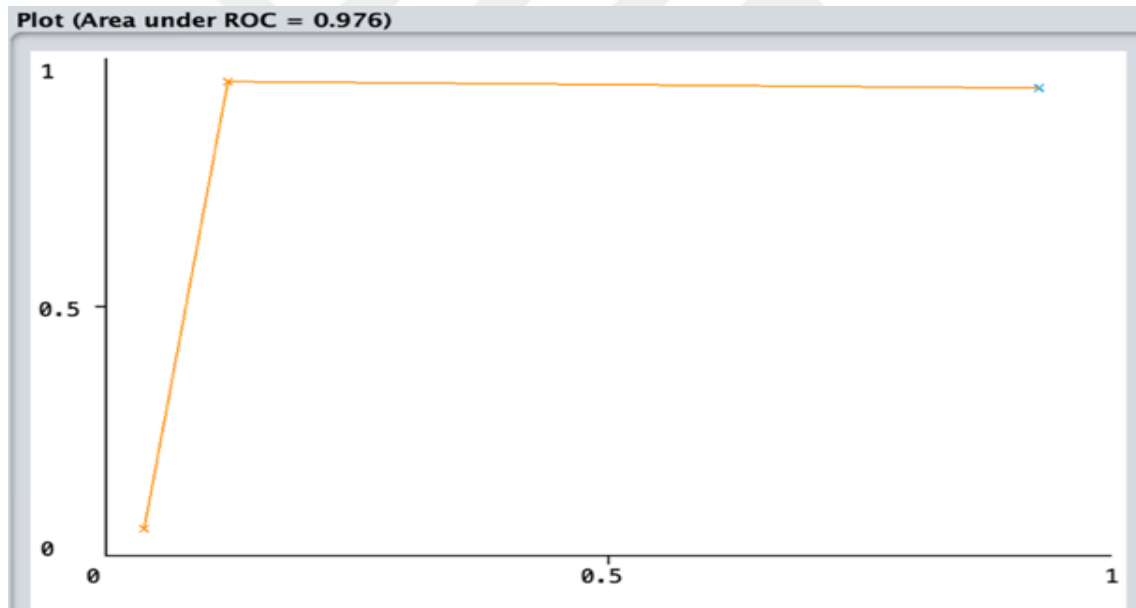
değeri 0.9713 çıkmıştır. Modelin örnekleri sınıflandırma başarısının iyi olduğu söylenebilir.

RBF için C parametresi değeri ikiyüz ($C=200$) seçilmiştir.

Çizelge 4.3 Göğüs kanseri veri seti RBF çekirdek fonksiyonu analiz sonucu

iyi huylu		kötü huylu		kappa istatistik değeri	doğruluk oranı
tp	tn	tp	tn		
355	2	9	203	0.9584	0.980668

Çizelge 4.3'te Radial basis fonksiyonu (RBF) yardımıyla veriler analiz edilmiştir. Doğruluk oranı 0.980668 çıkmıştır. Kappa istatistik değeri 0.9584 çıkmıştır. Bu da radial basis çekirdek fonksiyonu ile kurulan modelin iyi olduğu ve sınıflar arasındaki uyumun yüksek olduğu söylenebilir.



Şekil 4.3 Göğüs kanseri veri seti RBF ile analiz edilen örneklerle ait ROC eğrisi

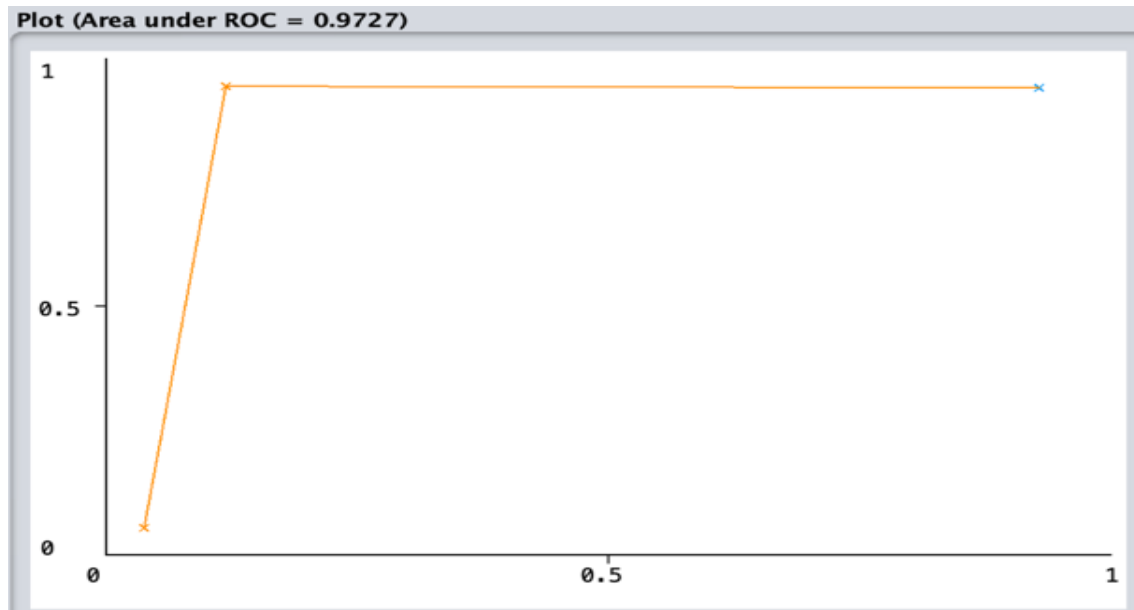
Şekilde 4.3'te iyi huylu ve kötü huylu örnekler RBF ile analiz edilmiş ve bu analiz sonucunda ROC eğrisi altında kalan alanın büyük olduğu görülmüştür. ROC değeri 0.976 çıkmıştır. Modelin örnekleri sınıflandırma başarısının iyi olduğu söylenebilir.

Sigmoid çekirdek fonksiyonu için C parametresi yirmidört ($C=24$).

Çizelge 4.4 Göğüs kanseri veri seti sigmoid çekirdek fonksiyonu analiz sonucu

iyi huylu		kötü huylu		kappa istatistik değeri	doğruluk oranı
tp	tn	tp	tn		
356	1	201	11	0.9545	0.97891

Çizelge 4.4'te Sigmoid çekirdek fonksiyonu yardımıyla veriler analiz edilmiştir. Doğruluk oranı 0.97891 çıkmıştır. Kappa istatistik değeri 0.9545 çıkmıştır. Bu da sigmoid çekirdek fonksiyonu ile kurulan modelin iyi olduğunu ve sınıflar arasındaki uyumun yüksek olduğu söylenebilir.



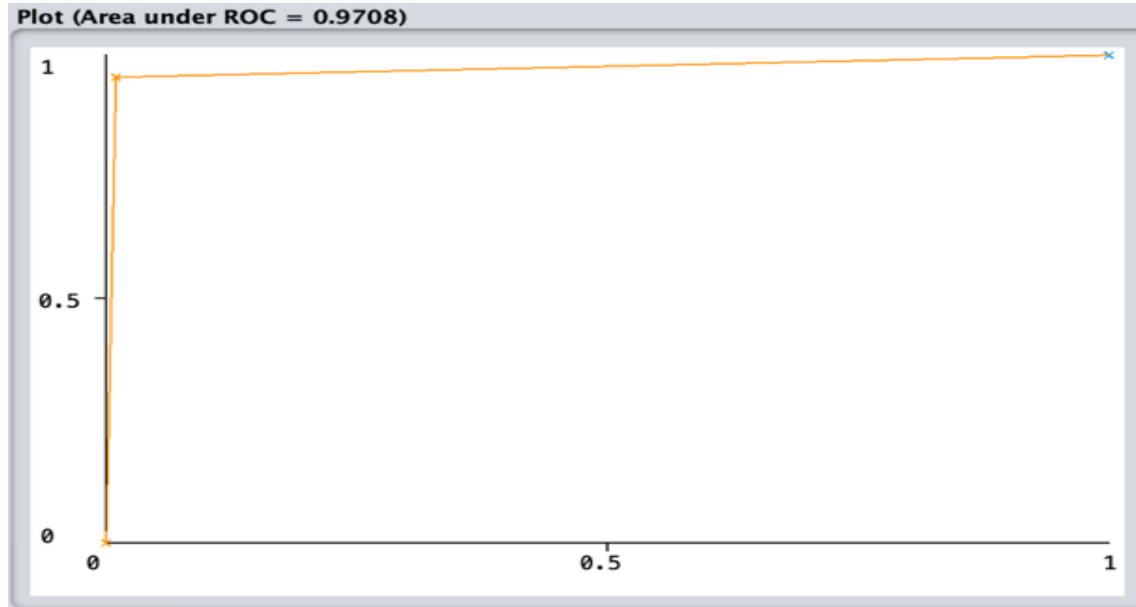
Şekil 4.4 Göğüs kanseri veri seti sigmoid çekirdek fonksiyonu ile analiz edilen örneklere ait ROC eğrisi

Şekilde 4.4'te iyi huylu ve kötü huylu örnekler sigmoid çekirdek fonksiyonu ile analiz edilmiş ve bu analiz sonucunda ROC eğrisi altında kalan alanın büyük olduğu görülmüştür. AUC değeri 0.9727 çıkmıştır. Modelin örnekleri sınıflandırma başarısının iyi olduğu söylenebilir.

Çizelge 4.5 Göğüs kanseri veri seti pearson VII işlevsel çekirdek fonksiyonu analiz sonucu

M		B		kappa istatistik değeri	doğruluk oranı
tp	tn	tp	tn		
202	10	353	4	0.9471	97.5395

Çizelge 4.5'te Pearson VII işlevsel çekirdek fonksiyonu yardımıyla analiz edilen verilerde doğruluk oranı 97.5395 çıkmıştır. Ayrıca kappa istatistik değerinin 0.9471 çıkması modelin güvenilir ve sınıflar arasındaki uyumun iyi olduğu söylenebilir.



Şekil 4.5 Göğüs kanseri veri seti pearson VII işlevsel çekirdek fonksiyonu ile analiz edilen örneklere ait ROC eğrisi

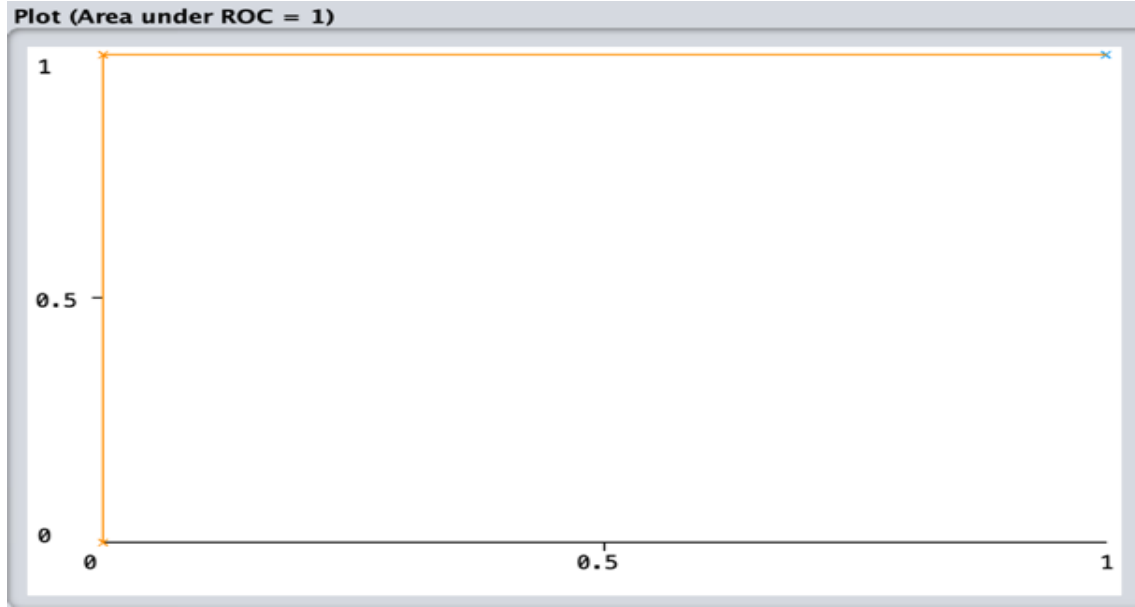
Şekil 4.5'te iyi huylu ve kötü örneklerine ait ROC eğrisinin altında kalan alanın yüksek olduğu görülmektedir. Ayrıca ROC değeri 0.9708 elde edilmiştir. Modelin örnekleri sınıflandırmada iyi olduğu söylenebilir.

4.2 Otizm Taraması Yetişkin Veri Seti Analiz Sonuçları

Çizelge 4.6 Otizm taraması yetişkin veri seti lineer çekirdek fonksiyonu analiz sonucu

No		Yes		kappa istatistik değeri	doğruluk oranı
tp	tn	tp	tn		
515	0	189	0	1	100

Çizelge 4.6’da Lineer çekirdek fonksiyonu yardımıyla analiz edilen verilerde doğruluk oranı 100 çıkmıştır. Ayrıca kappa istatistik değerinin 1 çıkması modelin güvenilir ve sınıflar arasındaki uyumun iyi olduğu söylenebilir.



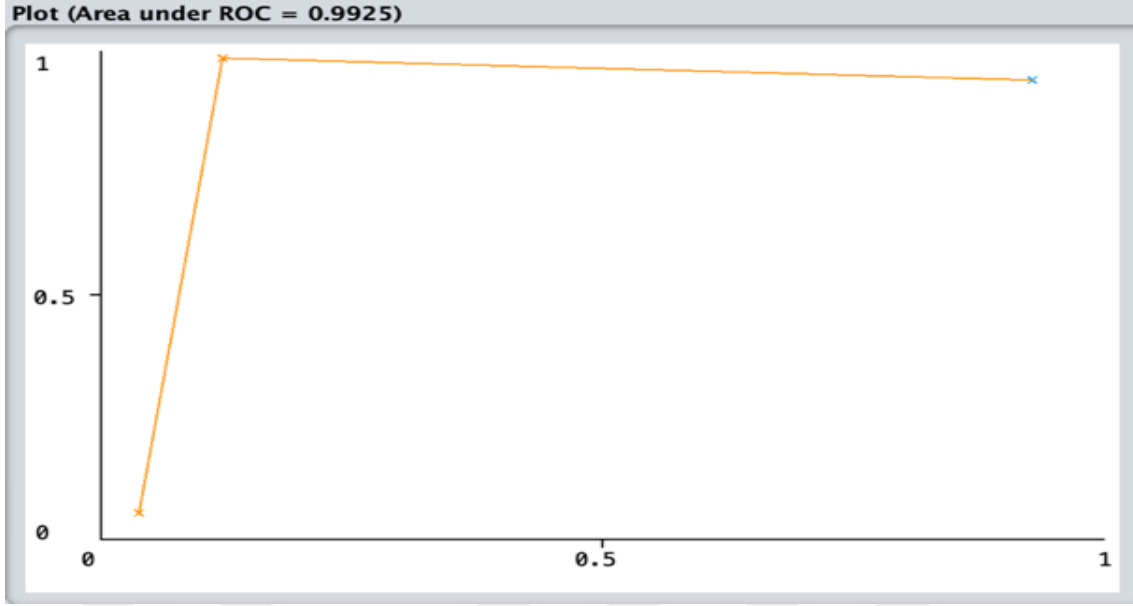
Şekil 4.6 Otizm taraması yetişkin veri seti lineer çekirdek fonksiyonu ile analiz edilen örnekler için ROC eğrisi

Şekil 4.6’da No ve Yes örneklerine ait ROC eğrisinin altında kalan alanın büyük olduğu görülmektedir. Ayrıca ROC değeri 1 elde edilmiştir. Modelin örnekleri sınıflandırmada başarılı olduğu söylenebilir.

Çizelge 4.7 Otizm taraması yetişkin veri seti polinom çekirdek fonksiyonu analiz sonucu

No		Yes		kappa istatistik değeri	doğruluk oranı
tp	tn	tp	tn		
510	05	188	1	0.9784	99.1477

Çizelge 4.7’de Polinom çekirdek fonksiyonu yardımıyla analiz edilen verilerde doğruluk oranı 99.1477 çıkmıştır. Ayrıca kappa istatistik değerinin 0.9784 çıkması modelin güvenilir ve sınıflar arasındaki uyumun iyi olduğu söylenebilir.



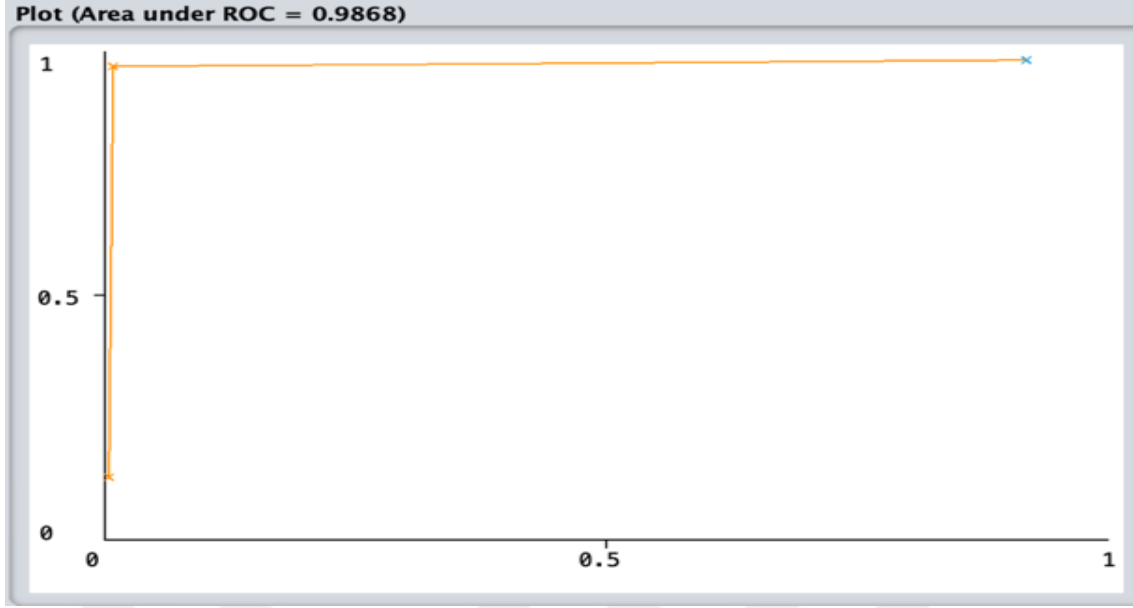
Şekil 4.7 Otizm taraması yetişkin veri seti polinom çekirdek fonksiyonu ile analiz edilen örneklere ait ROC eğrisi

Şekil 4.7’de No ve Yes örneklerine ait ROC eğrisinin altında kalan alanın büyük olduğu görülmektedir. Ayrıca ROC değeri 99.25 elde edilmiştir. Modelin örnekleri sınıflandırmada başarılı olduğu söylenebilir.

Çizelge 4.8 Otizm taraması yetişkin veri seti RBF çekirdek fonksiyonu analiz sonucu

No		Yes		kappa istatistik değeri doğruluk oranı	
tp	tn	tp	tn		
515	0	184	5	0.9818	99.2898

Çizelge 4.8’de RBF çekirdek fonksiyonu yardımıyla analiz edilen verilerde doğruluk oranı 99.2898 çıkmıştır. Ayrıca kappa istatistik değerinin 0.9818 çıkması modelin güvenilir ve sınıflar arasındaki uyumun iyi olduğu söylenebilir.



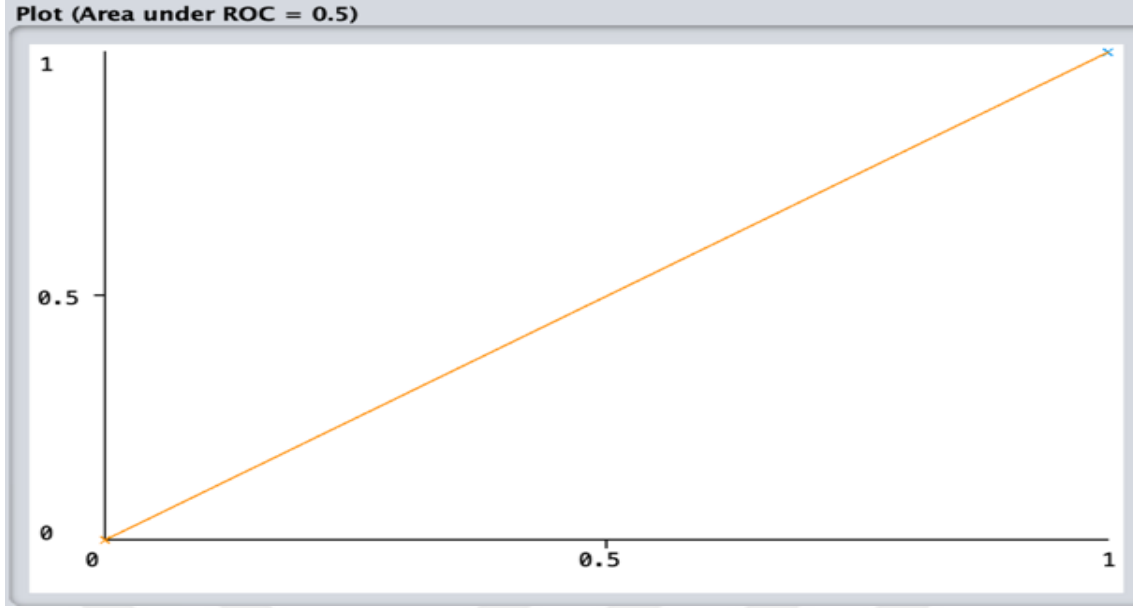
Şekil 4.8 Otizm taraması yetişkin veri seti RBF çekirdek fonksiyonu ile analiz edilen örneklere ait ROC eğrisi

Şekil 4.8’de No ve Yes örneklerine ait ROC eğrisinin altında kalan alanın büyük olduğu görülmektedir. Ayrıca ROC değeri 98.68 elde edilmiştir. Modelin örnekleri sınıflandırmada başarılı olduğu söylenebilir.

Çizelge 4.9 Otizm taraması yetişkin veri seti sigmoid çekirdek fonksiyonu analiz sonucu

No		Yes		kappa istatistik değeri	doğruluk oranı
tp	tn	tp	tn		
515	0	0	189	0	73.1534

Çizelge 4.9’da Sigmoid çekirdek fonksiyonu yardımıyla analiz edilen verilerde doğruluk oranı 73.1534 çıkmıştır. Ayrıca kappa istatistik değerinin 0 çıkması modelin güvenilirmez ve sınıflar arasındaki uyumun iyi olmadığı söylenebilir.



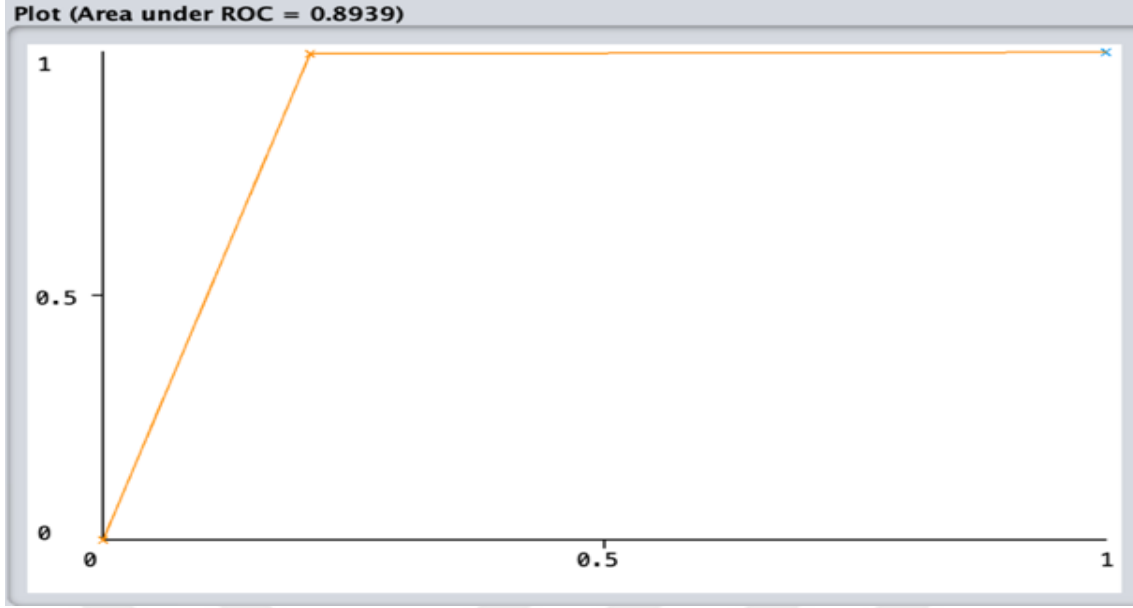
Şekil 4.9 Otizm taraması yetişkin veri seti sigmoid çekirdek fonksiyonu ile analiz edilen örneklere ait ROC eğrisi

Şekil 4.9’da No ve Yes örneklerine ait ROC eğrisinin altında kalan alanın küçük olduğu görülmektedir. Ayrıca ROC değeri 0.5 elde edilmiştir. Modelin örnekleri sınıflandırmada başarılı olduğu söylenebilir.

Çizelge 4.10 Otizm taraması yetişkin veri seti pearson VII işlevsel çekirdek fonksiyonu analiz sonucu

No		Yes		kappa istatistik değeri	doğruluk oranı
tp	tn	tp	tn		
512	03	150	39	0.8384	94.0341

Çizelge 4.10’da Pearson VII işlevsel çekirdek fonksiyonu yardımıyla analiz edilen verilerde doğruluk oranı 94.0341 çıkmıştır. Ayrıca kappa istatistik değerinin 0.8384 çıkması modelin güvenilir ve sınıflar arasındaki uyumun iyi olduğu söylenebilir.



řekil 4.10 Otizm taraması yetiřkin veri seti pearson VII iřlevsel çekirdek fonksiyonu ile analiz edilen örneklere ait ROC eğrisi

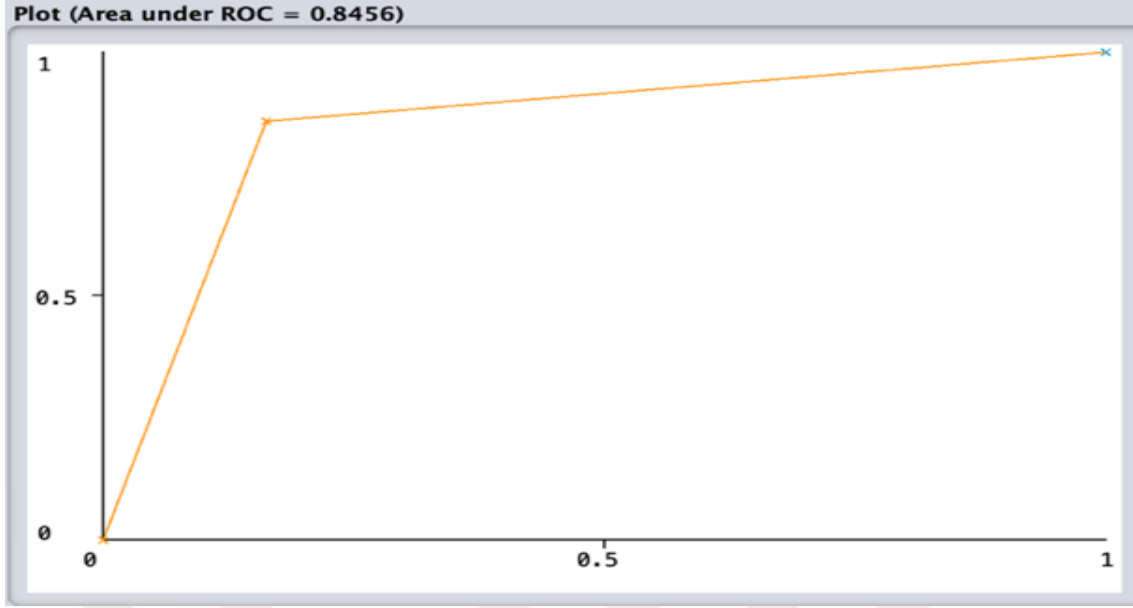
řekil 4.10’da No ve Yes örneklerine ait ROC eğrisinin altında kalan alanın büyük olduđu görölmektedir. Ayrıca ROC değeri 0.8939 elde edilmiştir. Modelin örnekleri sınıflandırmada başarılı olduđu söylenebilir.

4.3 Raisin Veri Seti Analiz Sonuçları

Çizelge 4.11 Raisin veri seti lineer çekirdek fonksiyonu analiz sonucu

Keçimen		Besni		kappa istatistik değeri	doğruluk oranı
tp	tn	tp	tn		
385	65	376	74	0.6911	84.5556

Çizelge 4.11’de Lineer çekirdek fonksiyonu yardımıyla analiz edilen verilerde doğruluk oranı 84.5556 çıkmıştır. Ayrıca kappa istatistik değerinin 0.6911 çıkması modelin güvenilir ve sınıflar arasındaki uyumun iyi olduđu söylenebilir.



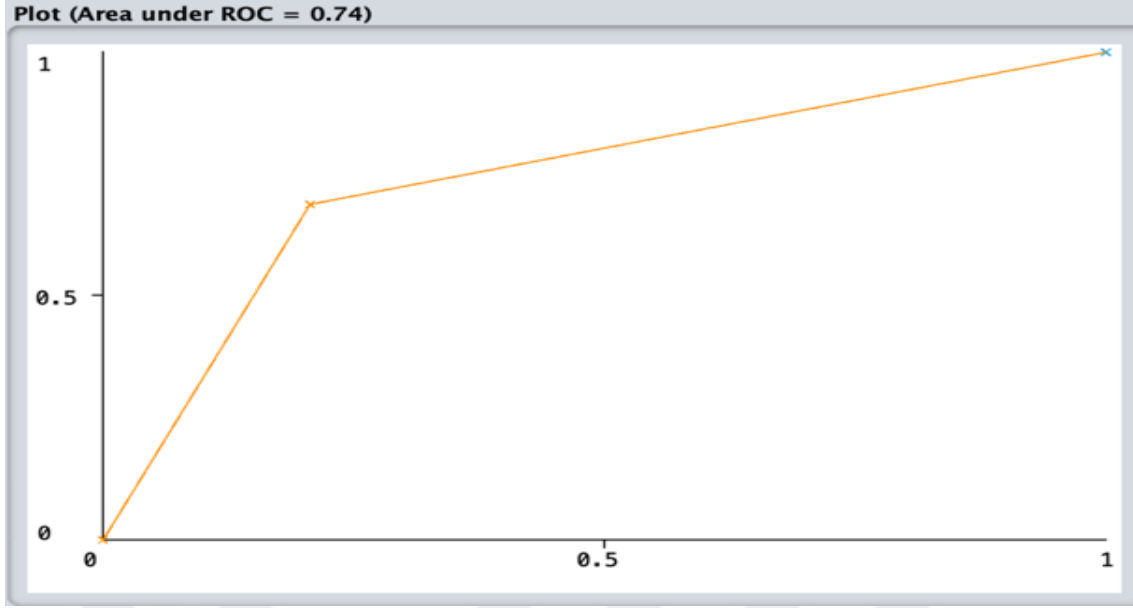
Şekil 4.11 Raisin veri seti lineer çekirdek fonksiyonu ile analiz edilen örneklere ait ROC eğrisi

Şekil 4.11’de Keçimen ve Besni örneklerine ait ROC eğrisinin altında kalan alanın büyük olduğu görülmektedir. Ayrıca ROC değeri 0.8456 elde edilmiştir. Modelin örnekleri sınıflandırmada başarılı olduğu söylenebilir.

Çizelge 4.12 Raisin veri seti polinom çekirdek fonksiyonu analiz sonucu

Keçimen		Besni		kappa istatistik değeri	doğruluk oranı
tp	tn	tp	tn		
309	141	357	93	0.48	74

Çizelge 4.12’de Polinom çekirdek fonksiyonu yardımıyla analiz edilen verilerde doğruluk oranı 74 çıkmıştır. Ayrıca kappa istatistik değerinin 0.48 çıkması modelin güvenilirmez ve sınıflar arasındaki uyumun iyi olmadığı söylenebilir.



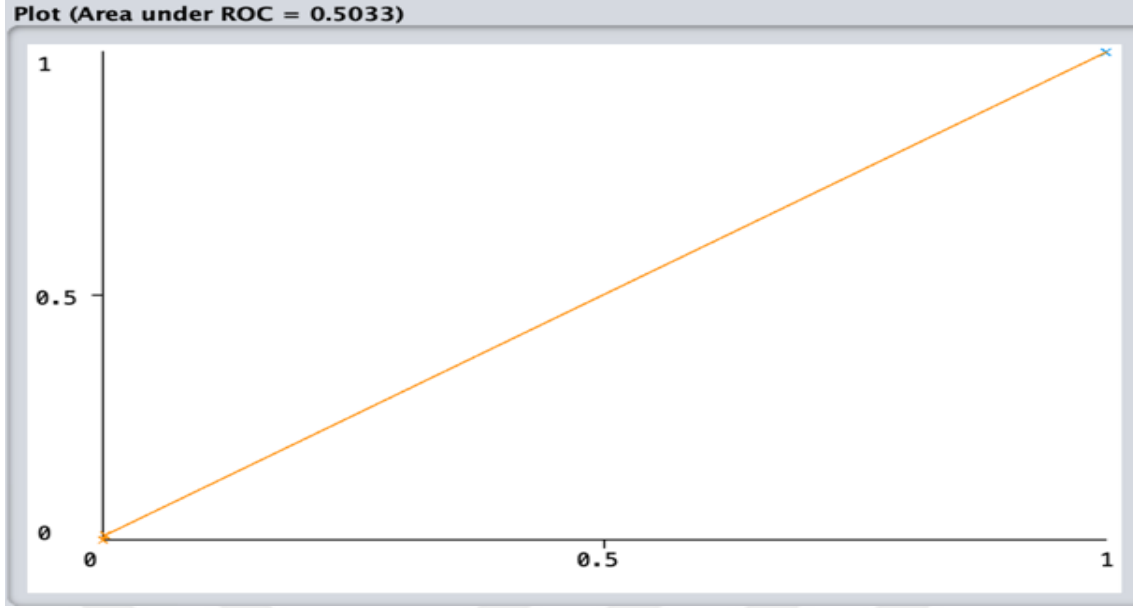
Şekil 4.12 Raisin veri seti polinom çekirdek fonksiyonu ile analiz edilen örneklere ait ROC eğrisi

Şekil 4.12’te Keçimen ve Besni örneklerine ait ROC eğrisinin altında kalan alanın büyük olduğu görülmektedir. Ayrıca ROC değeri 0.74 elde edilmiştir. Modelin örnekleri sınıflandırmada başarılı olduğu söylenebilir.

Çizelge 4.13 Raisin veri seti RBF çekirdek fonksiyonu analiz sonucu

Keçimen		Besni		kappa istatistik değeri	doğruluk oranı
tp	tn	tp	tn		
4	446	449	1	0.0067	50.3333

Çizelge 4.13’te RBF çekirdek fonksiyonu yardımıyla analiz edilen verilerde doğruluk oranı 50.3333 çıkmıştır. Ayrıca kappa istatistik değerinin 0.0067 çıkması modelin güvenilir ve sınıflar arasındaki uyumun iyi olmadığı söylenebilir.



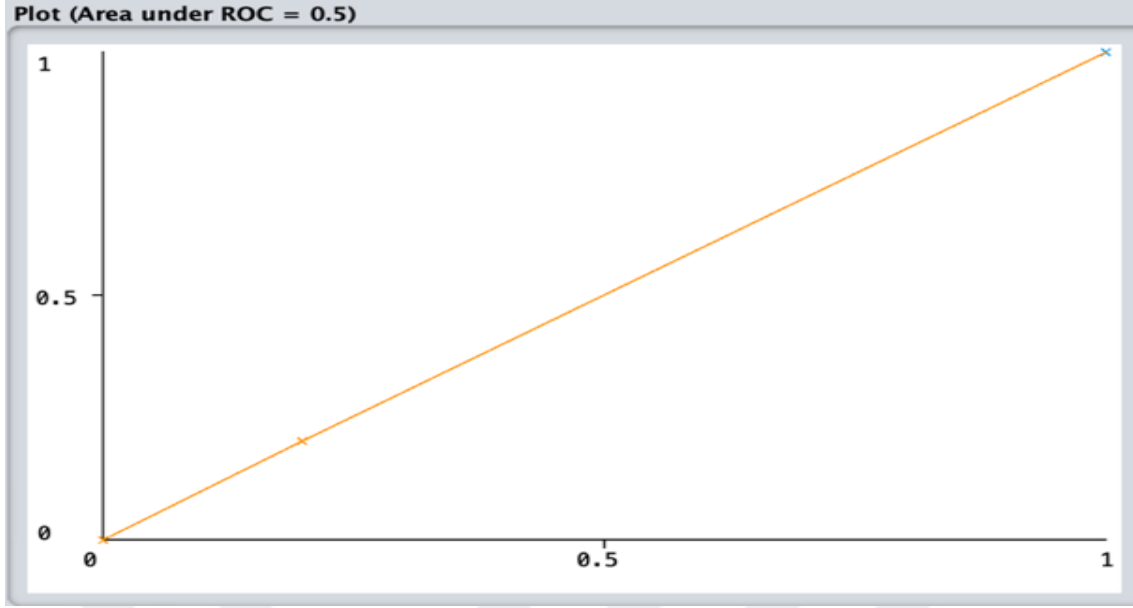
Şekil 4.13 Raisin veri seti RBF çekirdek fonksiyonu ile analiz edilen örneklere ait ROC eğrisi

Şekil 4.13'te Keçimen ve Besni örneklerine ait ROC eğrisinin altında kalan alanın küçük olduğu görülmektedir. Ayrıca ROC değeri 0.5033 elde edilmiştir. Modelin örnekleri sınıflandırmada başarısız olduğu söylenebilir.

Çizelge 4.14 Raisin veri seti sigmoid çekirdek fonksiyonu analiz sonucu

Keçimen		Besni		kappa istatistik değeri	doğruluk oranı
tp	tn	tp	tn		
90	360	360	90	0	50

Çizelge 4.14'te sigmoid çekirdek fonksiyonu yardımıyla analiz edilen verilerde doğruluk oranı 50 çıkmıştır. Ayrıca kappa istatistik değerinin 0 çıkması modelin güvenilirmez ve sınıflar arasındaki uyumun iyi olmadığı söylenebilir.



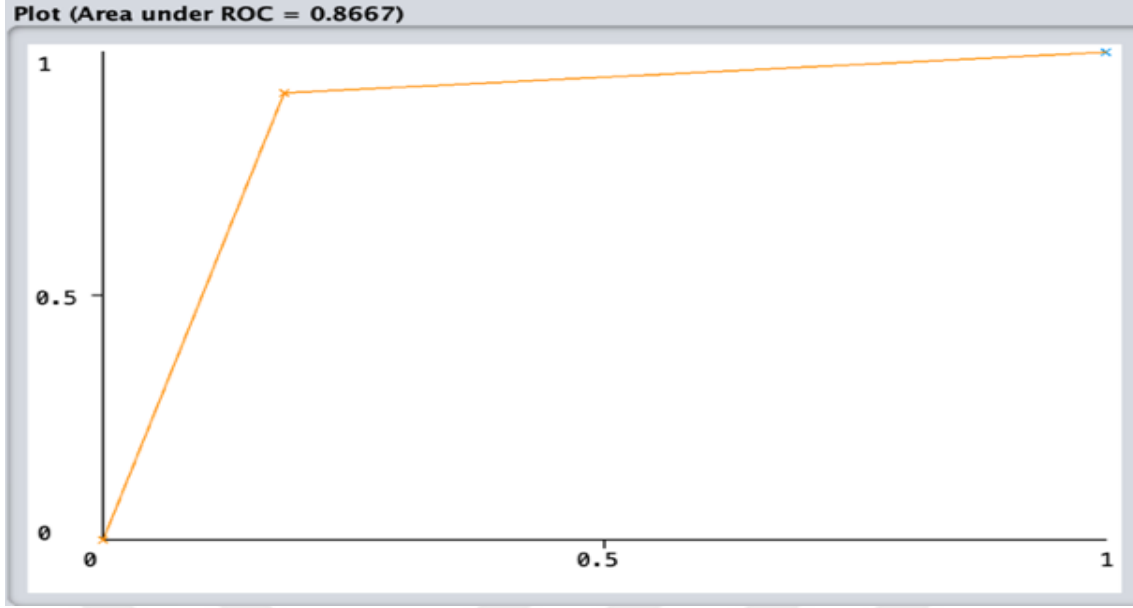
Şekil 4.14 Raisin veri seti Sigmoid çekirdek fonksiyonu ile analiz edilen örneklere ait ROC eğrisi

Şekil 4.14'te Keçimen ve Besni örneklerine ait ROC eğrisinin altında kalan alanın küçük olduğu görülmektedir. Ayrıca ROC değeri 0.5 elde edilmiştir. Modelin örnekleri sınıflandırmada başarısız olduğu söylenebilir.

Çizelge 4.15 Raisin veri seti pearson işlevsel VII çekirdek fonksiyonu analiz sonucu

Keçimen		Besni		kappa istatistik değeri	doğruluk oranı
tp	tn	tp	tn		
412	38	368	82	0.7333	86.6667

Çizelge 4.15'te Pearson işlevsel VII çekirdek fonksiyonu yardımıyla analiz edilen verilerde doğruluk oranı 86.6667 çıkmıştır. Ayrıca kappa istatistik değerinin 0.7333 çıkması modelin güvenilir ve sınıflar arasındaki uyumun iyi olduğu söylenebilir.



Şekil 4.15 Raisin veri seti pearson VII işlevsel çekirdek fonksiyonu ile analiz edilen örneklere ait ROC eğrisi

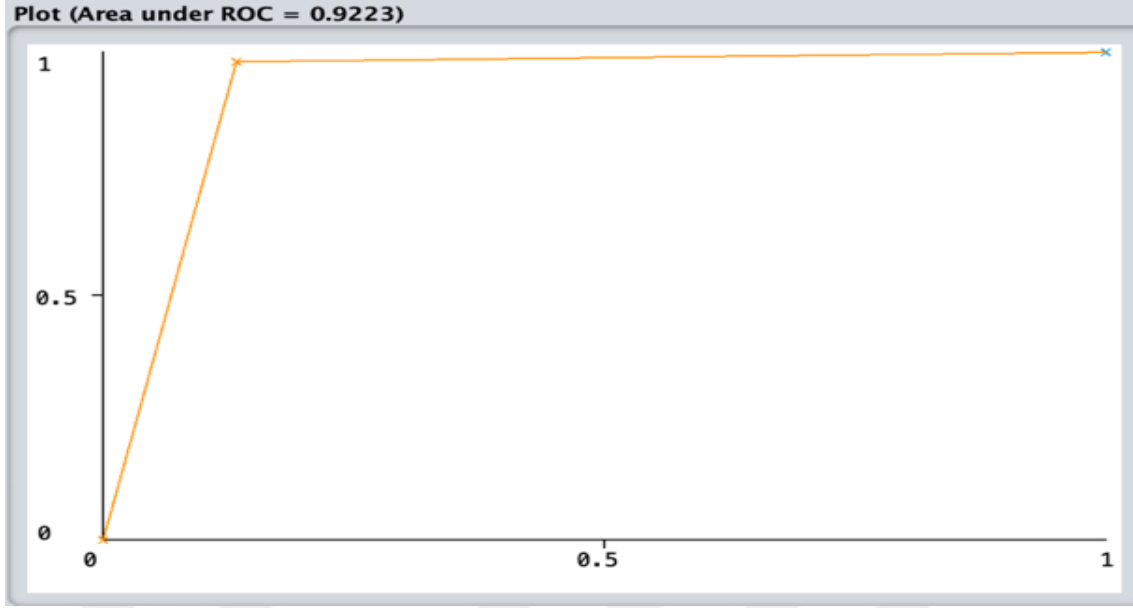
Şekil 4.15'te Keçimen ve Besni örneklerine ait ROC eğrisinin altında kalan alanın büyük olduğu görülmektedir. Ayrıca ROC değeri 0.8667 elde edilmiştir. Modelin örnekleri sınıflandırmada başarılı olduğu söylenebilir.

4.4 Voice Veri Seti Analiz Sonuçları

Çizelge 4.16 Voice veri seti lineer çekirdek fonksiyonu analiz sonucu

Male		Female		kappa istatistik değeri	doğruluk oranı
tp	tn	tp	tn		
1550	34	1372	212	0.8447	92.2348

Çizelge 4.16'da Lineer çekirdek fonksiyonu yardımıyla analiz edilen verilerde doğruluk oranı 92.2348 çıkmıştır. Ayrıca kappa istatistik değerinin 0.8447 çıkması modelin güvenilir ve sınıflar arasındaki uyumun iyi olduğu söylenebilir.



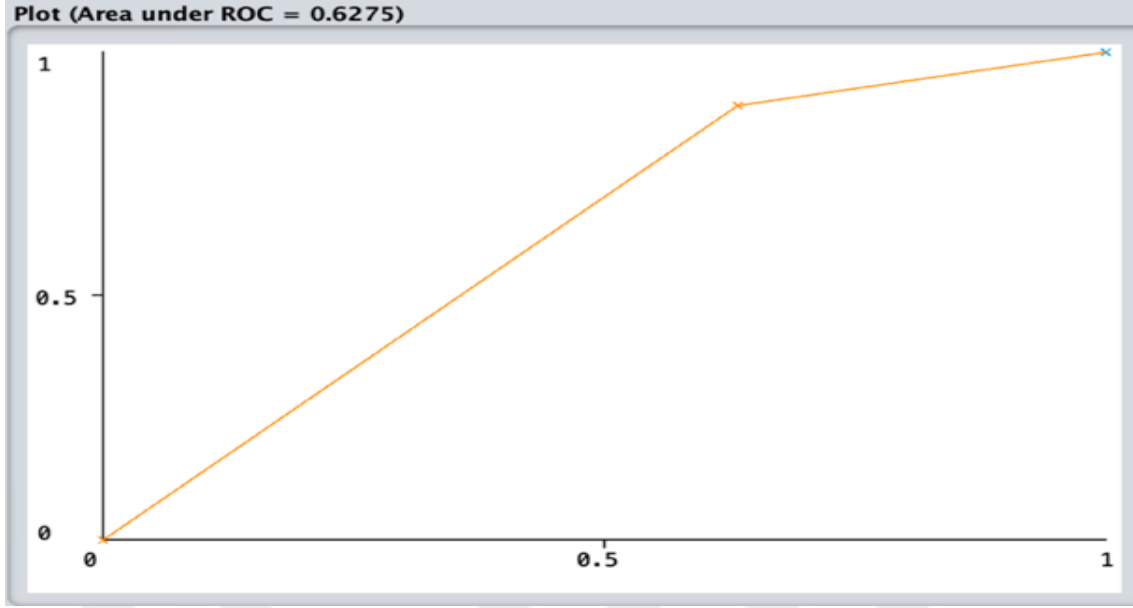
Şekil 4.16 Voice veri seti lineer çekirdek fonksiyonu ile analiz edilen örneklere ait ROC eğrisi

Şekil 4.16’da Male ve Female örneklerine ait ROC eğrisinin altında kalan alanın büyük olduğu görülmektedir. Ayrıca ROC değeri 0.9223 elde edilmiştir. Modelin örnekleri sınıflandırmada başarılı olduğu söylenebilir.

Çizelge 4.17 Voice veri seti polinom çekirdek fonksiyonu analiz sonucu

Male		Female		Kappa istatistik değeri	Doğruluk oranı
tp	tn	tp	tn		
1407	177	1003	581	0.2551	62.7525

Çizelge 4.17’de Polinom çekirdek fonksiyonu yardımıyla analiz edilen verilerde doğruluk oranı 62.7525 çıkmıştır. Ayrıca kappa istatistik değerinin 0.2551 çıkması modelin güvenilirmez ve sınıflar arasındaki uyumun iyi olmadığı söylenebilir.



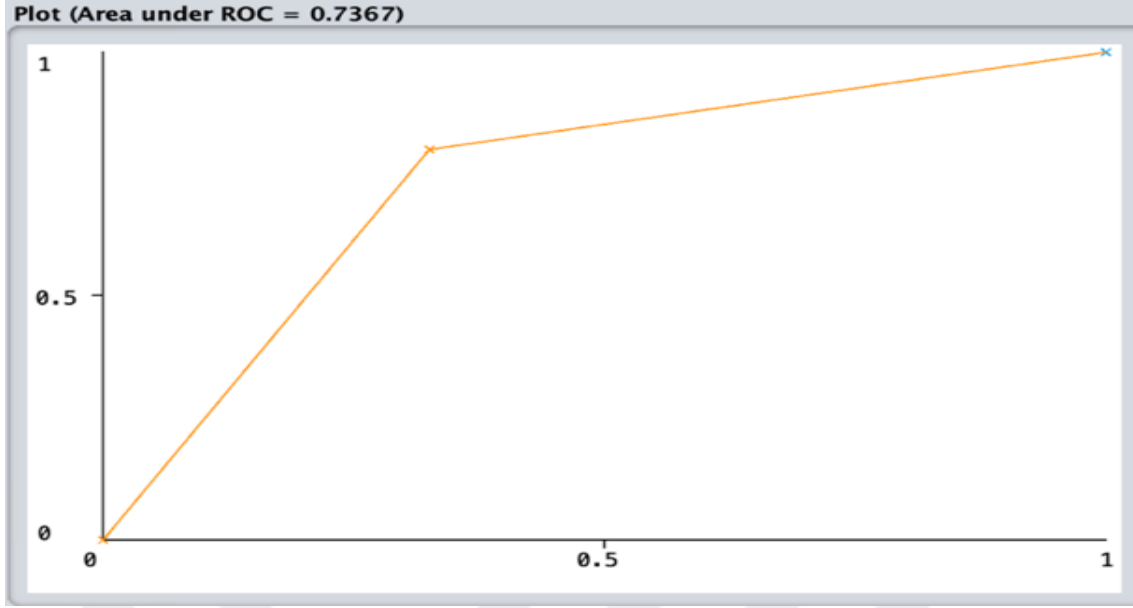
Şekil 4.17 Voice veri seti polinom çekirdek fonksiyonu ile analiz edilen örneklere ait ROC eğrisi

Şekil 4.17’de Male ve Female örneklerine ait ROC eğrisinin altında kalan alanın büyük olduğu görülmektedir. Ayrıca ROC değeri 0.6275 elde edilmiştir. Modelin örnekleri sınıflandırmada başarılı olduğu söylenebilir.

Çizelge 4.18 Voice veri seti RBF çekirdek fonksiyonu analiz sonucu

Male		Female		Kappa istatistik değeri	Doğruluk oranı
tp	tn	tp	tn		
1266	318	1068	516	0.4735	73.6742

Çizelge 4.18’de RBF çekirdek fonksiyonu yardımıyla analiz edilen verilerde doğruluk oranı 73.6742 çıkmıştır. Ayrıca kappa istatistik değerinin 0.4735 çıkması modelin güvenilir ve sınıflar arasındaki uyumun iyi olmadığı söylenebilir.



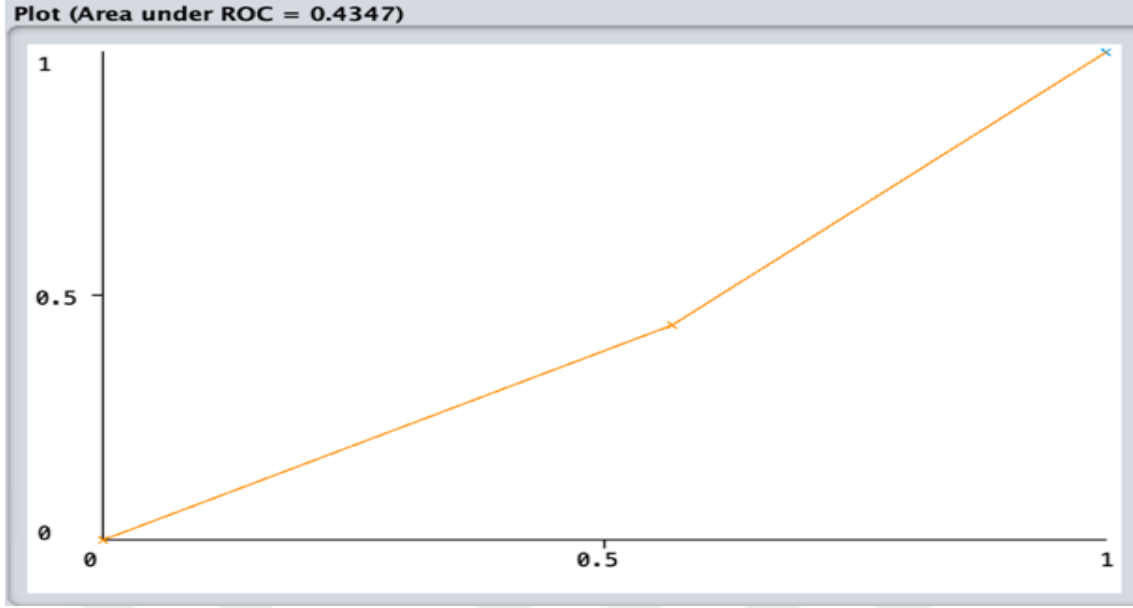
Şekil 4.18 Voice veri seti RBF çekirdek fonksiyonu ile analiz edilen örneklere ait ROC eğrisi

Şekil 4.18’de Male ve Female örneklerine ait ROC eğrisinin altında kalan alanın büyük olduğu görülmektedir. Ayrıca ROC değeri 0.7367 elde edilmiştir. Modelin örnekleri sınıflandırmada başarılı olduğu söylenebilir.

Çizelge 4.19 Voice veri seti sigmoid çekirdek fonksiyonu analiz sonucu

Male		Female		Kappa istatistik değeri	Doğruluk oranı
tp	tn	tp	tn		
693	891	684	900	-0.1307	43.4659

Çizelge 4.19’da Sigmoid çekirdek fonksiyonu yardımıyla analiz edilen verilerde doğruluk oranı 43.4659 çıkmıştır. Ayrıca kappa istatistik değerinin -0.1307 çıkması modelin güvenilirmez ve sınıflar arasındaki uyumun iyi olmadığı söylenebilir.



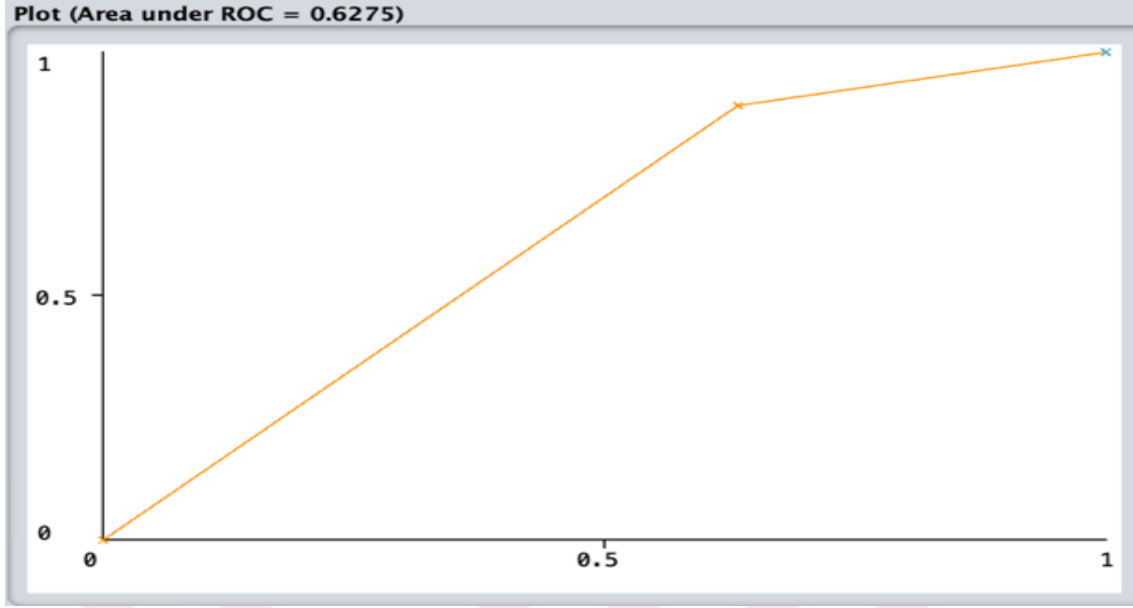
Şekil 4.19 Voice veri seti sigmoid çekirdek fonksiyonu ile analiz edilen örneklere ait ROC eğrisi

Şekil 4.19’de Male ve Female örneklerine ait ROC eğrisinin altında kalan alanın küçük olduğu görülmektedir. Ayrıca ROC değeri 0.4347 elde edilmiştir. Modelin örnekleri sınıflandırmada başarısız olduğu söylenebilir.

Çizelge 4.20 Voice veri seti pearson VII işlevsel çekirdek fonksiyonu analiz sonucu

Male		Female		Kappa istatistik değeri	Doğruluk oranı
tp	tn	tp	tn		
1407	177	1003	851	0.2551	62.7525

Çizelge 4.20’da Pearson VII işlevsel çekirdek fonksiyonu yardımıyla analiz edilen verilerde doğruluk oranı 62.7525 çıkmıştır. Ayrıca kappa istatistik değerinin 0.2551 çıkması modelin güvenilir ve sınıflar arasındaki uyumun iyi olmadığı söylenebilir.



Şekil 4.20 Voice veri seti Pearosn VII işlevsel çekirdek fonksiyonu ile analiz edilen örneklere ait ROC eğrisi

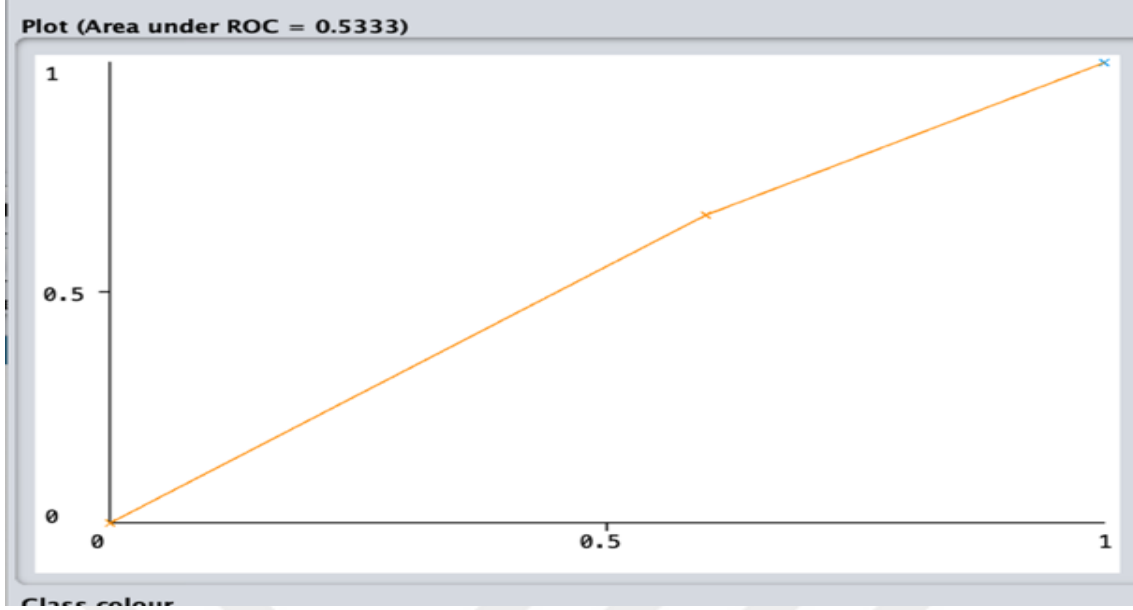
Şekil 4.20’de Male ve Female örneklerine ait ROC eğrisinin altında kalan alanın yeterince iyi olmadığı görülmektedir. Ayrıca ROC değeri 0.6275 elde edilmiştir. Modelin örnekleri sınıflandırmada çok iyi olmadığı söylenebilir.

4.5 Weather Veri Seti Analiz Sonuçları

Çizelge 4.21 Weather veri seti lineer çekirdek fonksiyonu analiz sonucu

yes		no		kappa istatistik değeri	doğruluk oranı
tp	tn	tp	tn		
3	3	3	2	0.0667	57.142

Çizelge 4.21’de lineer çekirdek fonksiyonu yardımıyla analiz edilen verilerde doğruluk oranı 57.142 çıkmıştır. Ayrıca kappa istatistik değerinin 0.0667 çıkması modelin güvenilir ve sınıflar arasındaki uyumun iyi olmadığı söylenebilir.



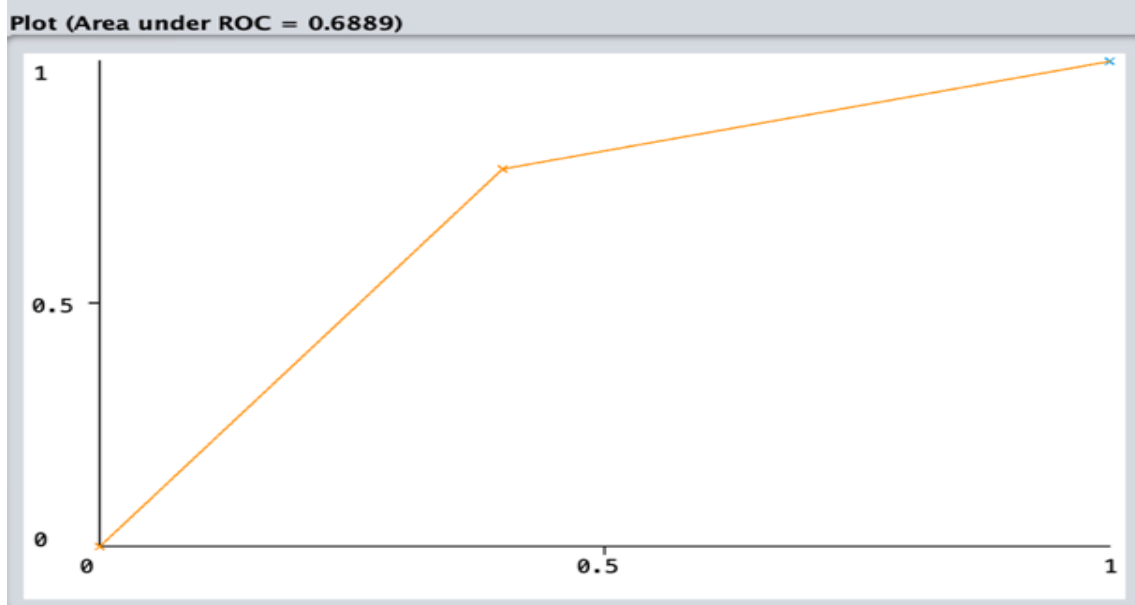
Şekil 4.21 Weather veri seti lineer çekirdek fonksiyonu ile analiz edilen örneklere ait ROC eğrisi

Şekil 4.21’de yes ve no örneklere ait ROC eğrisinin altında kalan alanın büyük olmadığı görülmektedir. Ayrıca ROC değeri 0.5333 elde edilmiştir. Modelin örnekleri sınıflandırma başarısının iyi olduğu söylenemez.

Çizelge 4.22 Weather veri seti polinom çekirdek fonksiyonu analiz sonucu

yes		no		kappa istatistik değeri	doğruluk oranı
tp	tn	tp	tn		
7	2	2	3	0.3778	71.4286

Çizelge 4.22’de polinom çekirdek fonksiyonu yardımıyla analiz edilen verilerde doğruluk oranı 71.42 çıkmıştır. Ayrıca kappa istatistik değerinin 0.3778 çıkması modelin güvenilir ve sınıflar arasındaki uyumun iyi olmadığı söylenebilir.



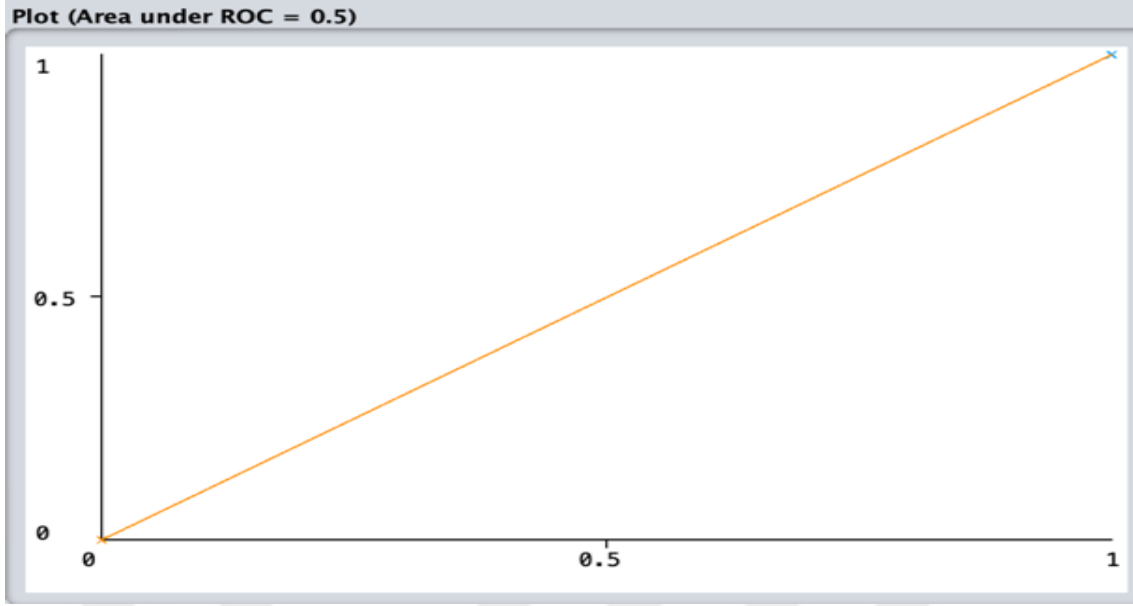
Şekil 4.22 Weather veri seti Polinom çekirdek fonksiyonu ile analiz edilen örneklere ait ROC eğrisi

Şekil 4.22’de yes ve no örneklerine ait ROC eğrisinin altında kalan alanın büyük olduğu görülmektedir. Ayrıca ROC değeri 0.6889 elde edilmiştir. Modelin örnekleri sınıflandırma başarısının iyi olduğu söylenebilir.

Çizelge 4.23 Weather veri seti RBF ve sigmoid çekirdek fonksiyonu analiz sonucu

yes		no		kappa istatistik değeri	doğruluk oranı
tp	tn	tp	tn		
9	0	0	5	0	54.2857

Çizelge 4.23’de RBF ve Sigmoid çekirdek fonksiyonu yardımıyla analiz edilen verilerde doğruluk oranı 54.2857 çıkmıştır. Ayrıca kappa istatistik değerinin 0 çıkması modelin güvenilmez ve sınıflar arasındaki uyumun iyi olmadığı söylenebilir



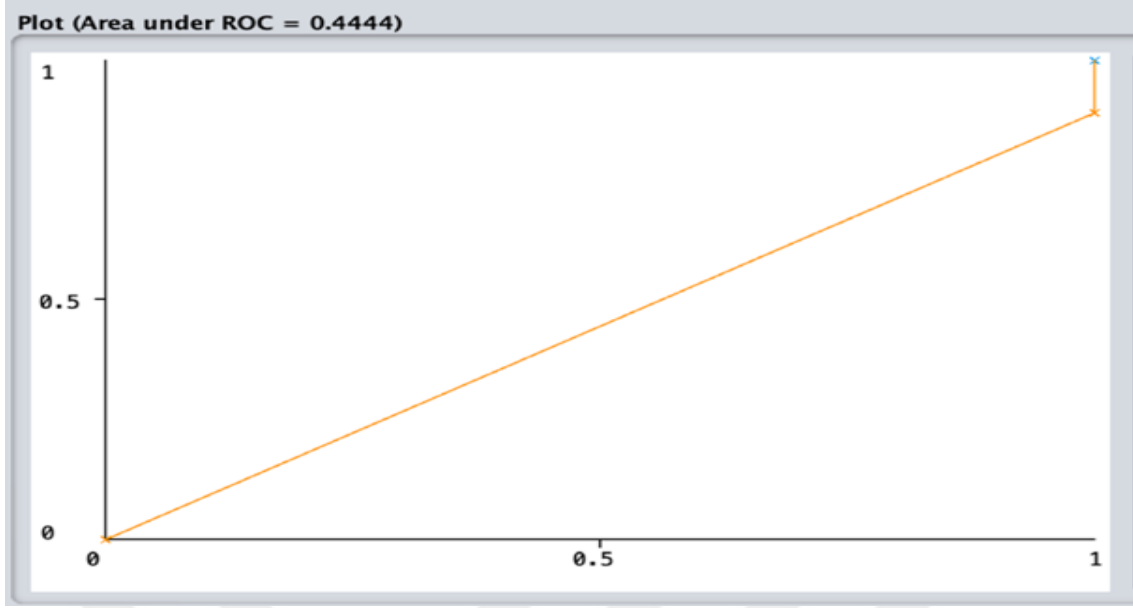
Şekil 4.23 Weather veri seti RBF ve sigmoid çekirdek fonksiyonu ile analiz edilen örneklere ait ROC eğrisi

Şekil 4.23'te yes ve no örneklerine ait ROC eğrisinin altında kalan alanın küçük olduğu görülmektedir. Ayrıca ROC değeri 0.5 elde edilmiştir. Modelin örnekleri sınıflandırma başarısının iyi olmadığı söylenebilir.

Çizelge 4.24 Weather veri seti pearson VII İşlevsel çekirdek fonksiyonu analiz sonucu

yes		no		kappa istatistik değeri	doğruluk oranı
tp	tn	tp	tn		
8	01	0	5	-0.1351	57.1429

Çizelge 4.24'te Pearson VII işlevsel çekirdek fonksiyonu yardımıyla analiz edilen verilerde doğruluk oranı 57.1429 çıkmıştır. Ayrıca kappa istatistik değerinin -0.1351 çıkması modelin güvenilirmez ve sınıflar arasındaki uyumun iyi olmadığı söylenebilir.



Şekil 4.24 Weather veri seti pearson VII işlevsel çekirdek fonksiyonu ile analiz edilen örneklere ait ROC eğrisi

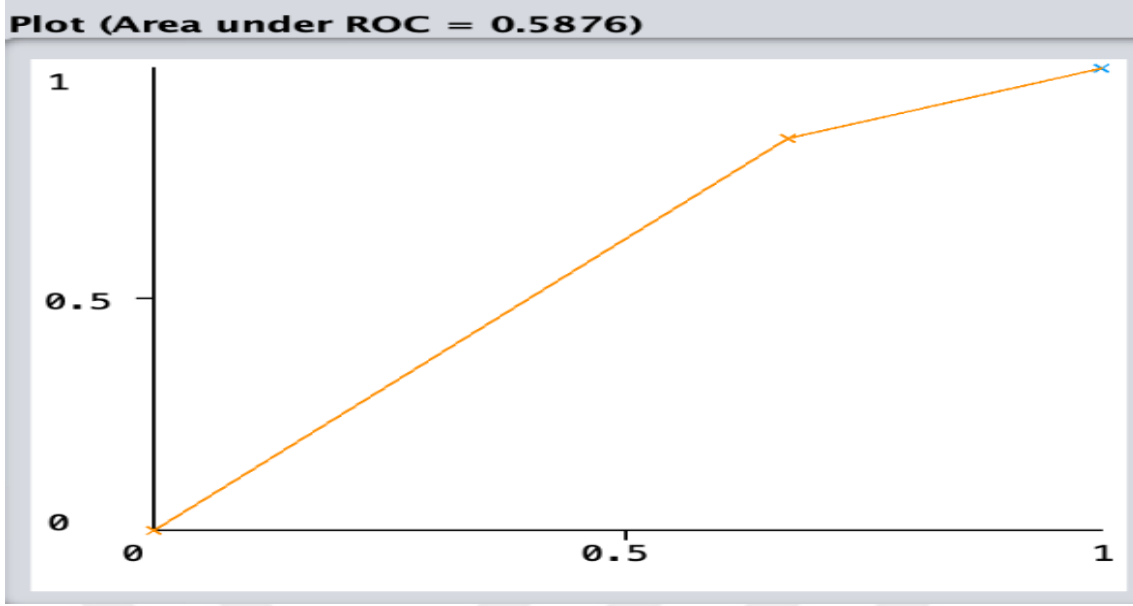
Şekil 4.24'te yes ve no örneklerine ait ROC eğrisinin altında kalan alanın küçük olduğu görülmektedir. Ayrıca ROC değeri 0.44 elde edilmiştir. Modelin örnekleri sınıflandırma başarısının iyi olmadığı söylenebilir.

4.6 Breast Cancer Veri Seti Analiz Sonuçları

Çizelge 4.25 Breast cancer veri seti lineer çekirdek fonksiyonu analiz sonucu

no recurrence-event		recurrence event		kappa istatistik değeri	doğruluk oranı
tp	tn	tp	tn		
170	31	57	28	0.19	69.2308

Çizelge 4.25'te lineer çekirdek fonksiyonu yardımıyla analiz edilen verilerde doğruluk oranı 69.2308 çıkmıştır. Ayrıca kappa istatistik değerinin 0.19 çıkması modelin güvenilir ve sınıflar arasındaki uyumun iyi olmadığı söylenebilir.



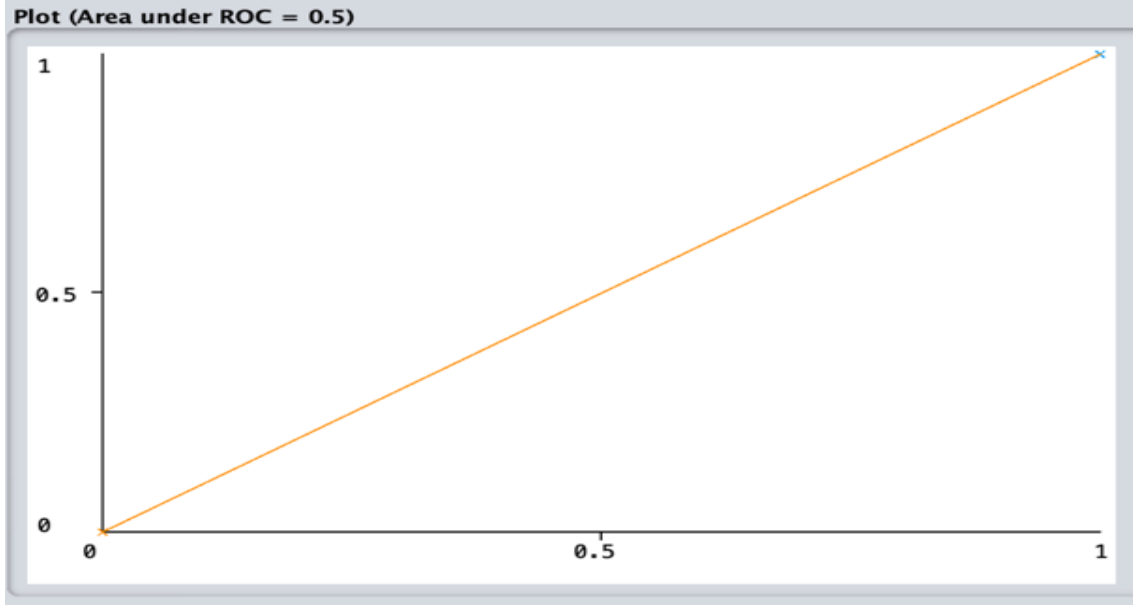
Şekil 4.25 Breast cancer veri seti lineer çekirdek fonksiyonu ile analiz edilen örneklere ait ROC eğrisi

Şekil 4.25'te no recurrence-event recurrence event örneklerine ait ROC eğrisinin altında kalan alanın büyük olduğu görülmektedir. Ayrıca ROC değeri 0.5876 elde edilmiştir. Modelin örnekleri sınıflandırma başarısının iyi olduğu söylenebilir.

Çizelge 4.26 Breast cancer veri seti polinom çekirdek fonksiyonu analiz sonucu

no recurrence-event		recurrence event		kappa istatistik değeri	doğruluk oranı
tp	tn	tp	tn		
201	0	0	85	0	70.2797

Çizelge 4.26'da Polinom çekirdek fonksiyonu yardımıyla analiz edilen verilerde doğruluk oranı 70.2797 çıkmıştır. Ayrıca kappa istatistik değerinin 0 çıkması modelin güvenilir ve sınıflar arasındaki uyumun iyi olmadığı söylenebilir.



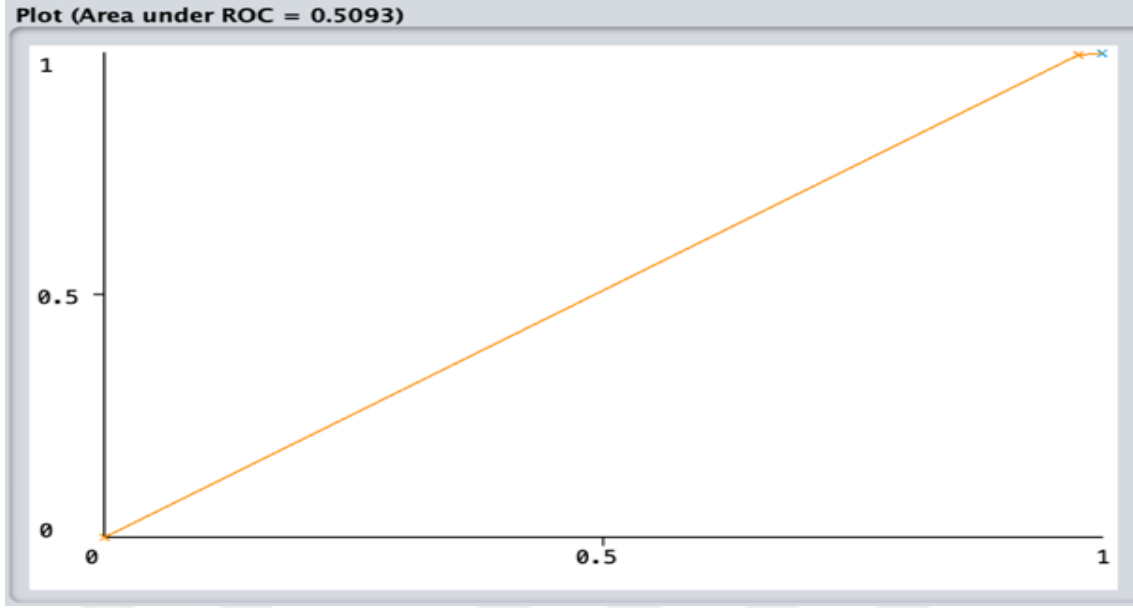
Şekil 4.26 Breast cancer veri seti polinom çekirdek fonksiyonu ile analiz edilen örneklere ait ROC eğrisi

Şekil 4.26’da no recurrence-event recurrence event örneklerine ait ROC eğrisinin altında kalan alanın küçük olduğu görülmektedir. Ayrıca ROC değeri 0.5 elde edilmiştir. Modelin örnekleri sınıflandırmada başarısız olduğu söylenebilir.

Çizelge 4.27 Breast cancer veri seti RBF çekirdek fonksiyonu analiz sonucu

no recurrence-event		recurrence event		kappa istatistik değeri	doğruluk oranı
tp	tn	tp	tn		
200	1	2	83	0.257	70.6294

Çizelge 4.27’de RBF çekirdek fonksiyonu yardımıyla analiz edilen verilerde doğruluk oranı 70.6294 çıkmıştır. Ayrıca kappa istatistik değerinin 0.257 çıkması modelin güvenilirmez ve sınıflar arasındaki uyumun iyi olmadığı söylenebilir.



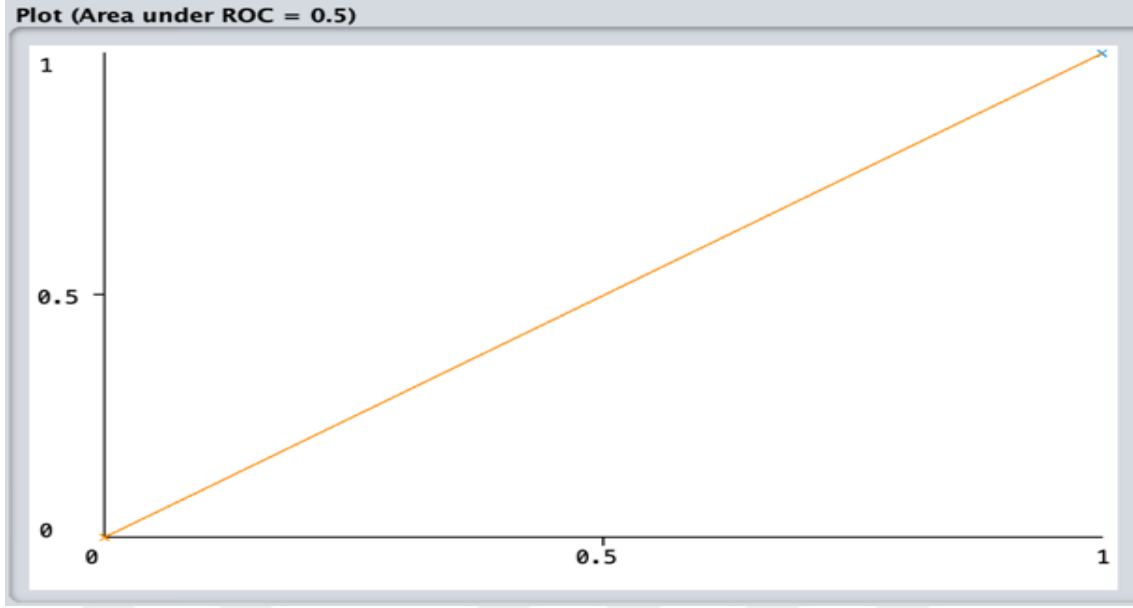
Şekil 4.27 Breast cancer veri seti RBF çekirdek fonksiyonu ile analiz edilen örneklere ait ROC eğrisi

Şekil 4.27’de no recurrence-event recurrence event örneklerine ait ROC eğrisinin altında kalan alanın küçük olduğu görülmektedir. Ayrıca ROC değeri 0.5093 elde edilmiştir. Modelin örnekleri sınıflandırmada başarısız olduğu söylenebilir.

Çizelge 4.28 Breast cancer veri seti sigmoid çekirdek fonksiyonu analiz sonucu

no recurrence-event		recurrence event		kappa istatistik değeri	doğruluk oranı
tp	tn	tp	tn		
200	1	2	83	0	70.2797

Çizelge 4.28’de Sigmoid çekirdek fonksiyonu yardımıyla analiz edilen verilerde doğruluk oranı 70.2797 çıkmıştır. Ayrıca kappa istatistik değerinin 0 çıkması modelin güvenilir ve sınıflar arasındaki uyumun iyi olmadığı söylenebilir.



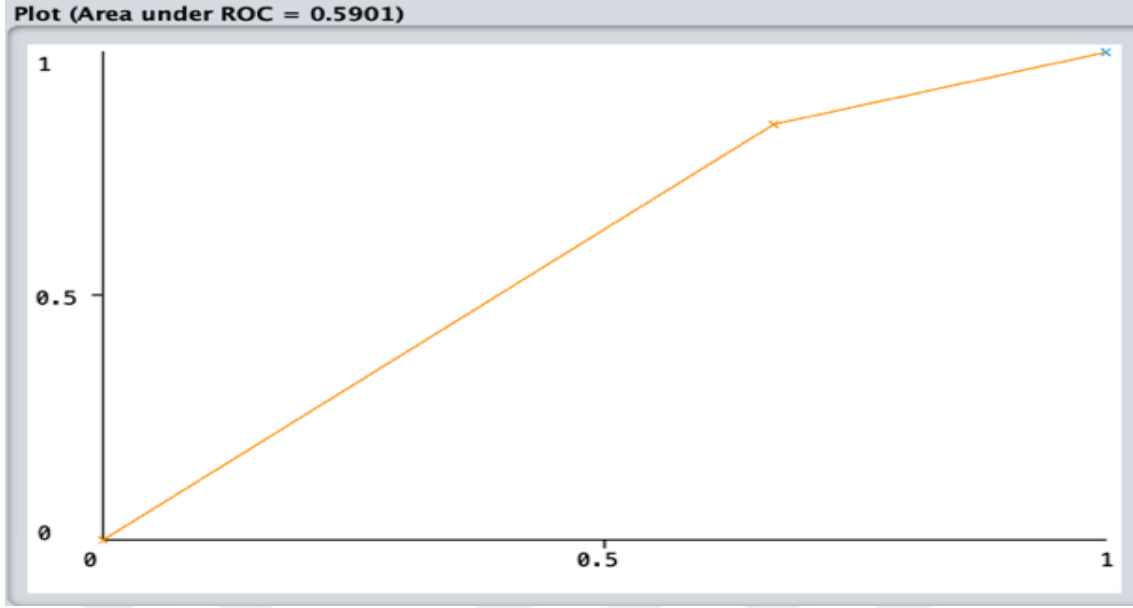
Şekil 4.28 Breast cancer veri seti sigmoid çekirdek fonksiyonu ile analiz edilen örneklere ait ROC eğrisi

Şekil 4.28’de no recurrence-event recurrence event örneklerine ait ROC eğrisinin altında kalan alanın küçük olduğu görülmektedir. Ayrıca ROC değeri 0.5 elde edilmiştir. Modelin örnekleri sınıflandırmada başarısız olduğu söylenebilir.

Çizelge 4.29 Breast cancer veri seti Pearson VII İşsel çekirdek fonksiyonu analiz sonucu

no recurrence-event		recurrence event		kappa istatistik değeri	doğruluk oranı
tp	tn	tp	tn		
171	30	28	57	0.1983	69.5804

Çizelge 4.29’da Pearson VII işlevsel çekirdek fonksiyonu yardımıyla analiz edilen verilerde doğruluk oranı 69.5804 çıkmıştır. Ayrıca kappa istatistik değerinin 0.1983 çıkması modelin güvenilir ve sınıflar arasındaki uyumun iyi olmadığı söylenebilir.



Şekil 4.29 Breast cancer veri seti pearson VII işlevsel çekirdek fonksiyonu ile analiz edilen örneklere ait ROC eğrisi

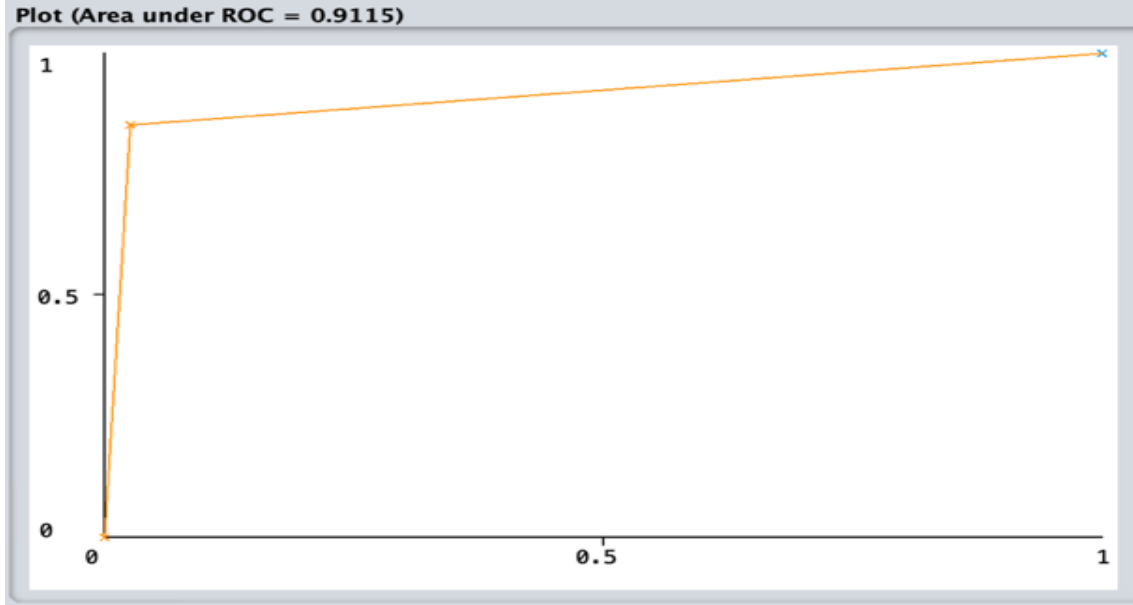
Şekil 4.29’da no recurrence-event recurrence event örneklerine ait ROC eğrisinin altında kalan alanın küçük olduğu görülmektedir. Ayrıca ROC değeri 0.5901 elde edilmiştir. Modelin örnekleri sınıflandırmada başarısız olduğu söylenebilir.

4.7 Labor neg Veri Seti Analiz Sonuçları

Çizelge 4.30 Labor neg veri seti lineer ve RBF çekirdek fonksiyonu analiz sonucu

bad		god		kappa istatistik değeri	doğruluk oranı
tp	tn	tp	tn		
17	3	36	1	0.8423	92.9825

Çizelge 4.30’da lineer ve RBF çekirdek fonksiyonu yardımıyla analiz edilen verilerde doğruluk oranı 92.9825 çıkmıştır. Ayrıca kappa istatistik değerinin 0.8423 çıkması modelin güvenilir ve sınıflar arasındaki uyumun iyi olduğu söylenebilir.



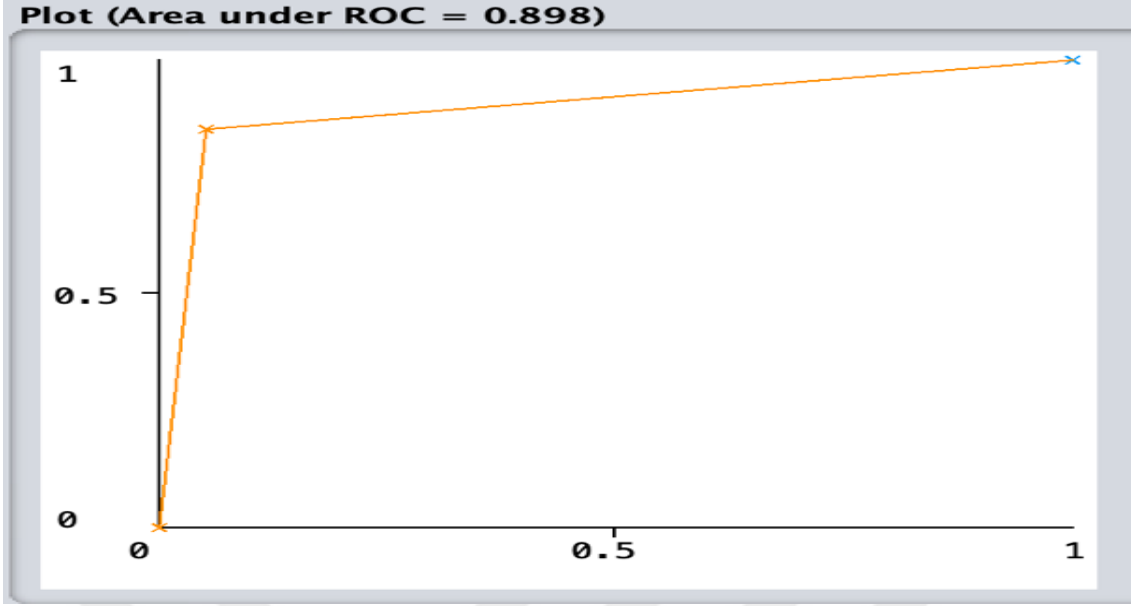
Şekil 4.30 Labor neg veri seti lineer ve RBF çekirdek fonksiyonu ile analiz edilen örneklerle ait ROC eğrisi

Şekil 4.30’da bad ve god örneklerine ait ROC eğrisinin altında kalan alanın büyük olduğu görülmektedir. Ayrıca ROC değeri 0.9115 elde edilmiştir. Modelin örnekleri sınıflandırmada başarılı olduğu söylenebilir.

Çizelge 4.31 Labor neg veri seti polinom çekirdek fonksiyonu analiz sonucu

bad		god		kappa istatistik değeri	doğruluk oranı
tp	tn	tp	tn		
17	3	35	2	0.8052	91.2281

Çizelge 4.31’de Polinom çekirdek fonksiyonu yardımıyla analiz edilen verilerde doğruluk oranı 91.2282 çıkmıştır. Ayrıca kappa istatistik değerinin 0.8052 çıkması modelin güvenilir ve sınıflar arasındaki uyumun iyi olduğu söylenebilir.



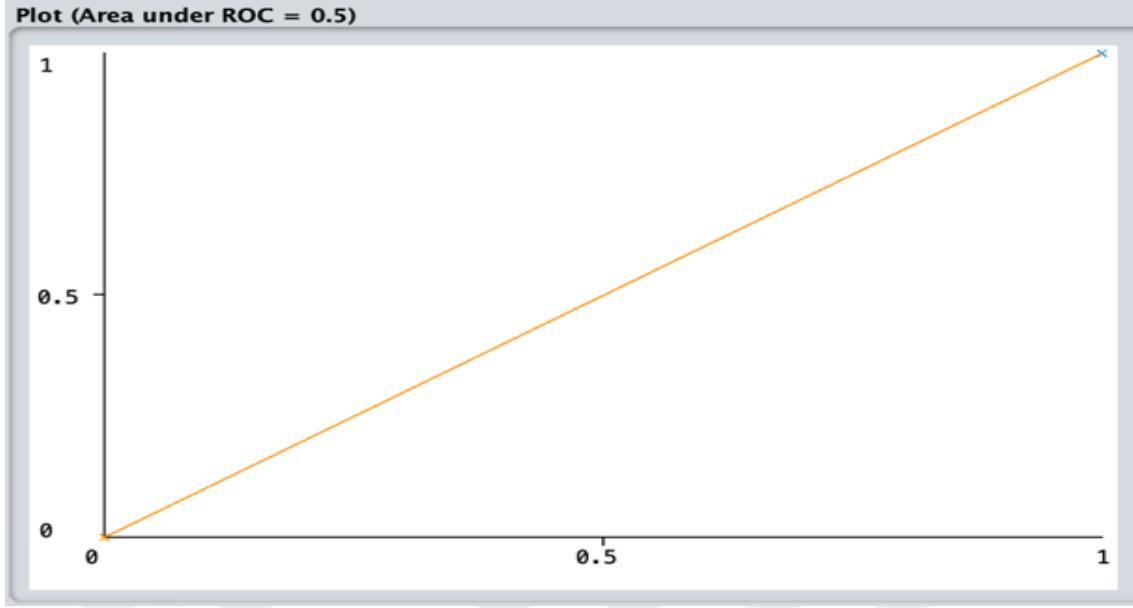
Şekil 4.31 Labor neg veri seti polinom çekirdek fonksiyonu ile analiz edilen örneklere ait ROC eğrisi

Şekil 4.31’de bad ve god örneklerine ait ROC eğrisinin altında kalan alanın büyük olduğu görülmektedir. Ayrıca ROC değeri 0.898 elde edilmiştir. Modelin örnekleri sınıflandırmada başarılı olduğu söylenebilir.

Çizelge 4.32 Labor neg veri seti sigmoid çekirdek fonksiyonu analiz sonucu

bad		god		kappa istatistik değeri	doğruluk oranı
tp	tn	tp	tn		
0	20	37	0	0.35079	64.9123

Çizelge 4.32’de sigmoid çekirdek fonksiyonu yardımıyla analiz edilen verilerde doğruluk oranı 64.9123 çıkmıştır. Ayrıca kappa istatistik değerinin 0.3509 çıkması modelin güvenilirmez ve sınıflar arasındaki uyumun iyi olmadığı söylenebilir.



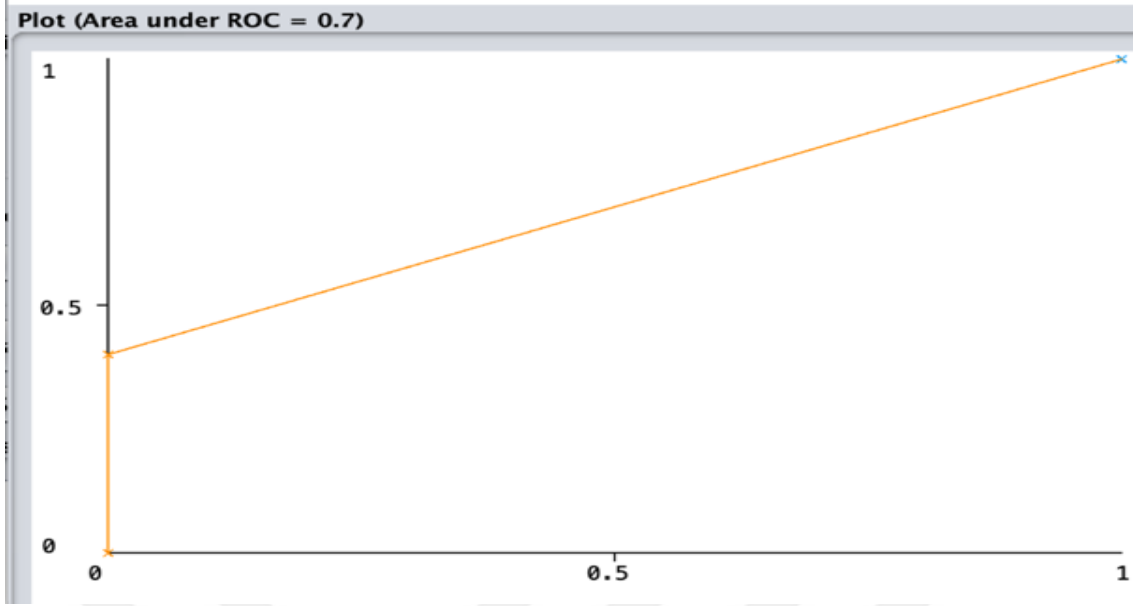
Şekil 4.32 Labor neg veri seti sigmoid çekirdek fonksiyonu ile analiz edilen örneklere ait ROC eğrisi

Şekil 4.32’de bad ve god örneklerine ait ROC eğrisinin altında kalan alanın büyük olduğu görülmektedir. Ayrıca ROC değeri 0.898 elde edilmiştir. Modelin örnekleri sınıflandırmada başarılı olduğu söylenebilir.

Çizelge 4.33 Labor neg veri seti Pearson VII işlevsel çekirdek fonksiyonu analiz sonucu

bad		god		kappa istatistik değeri	doğruluk oranı
tp	tn	tp	tn		
8	12	37	0	0.4639	78.9474

Çizelge 4.33’te Pearson VII işlevsel çekirdek fonksiyonu yardımıyla analiz edilen verilerde doğruluk oranı 78.9474 çıkmıştır. Ayrıca kappa istatistik değerinin 0.4639 çıkması modelin güvenilir ve sınıflar arasındaki uyumun iyi olmadığı söylenebilir.



Şekil 4.33 Labor neg veri seti pearson VII işlevsel çekirdek fonksiyonu ile analiz edilen örneklere ait ROC eğrisi

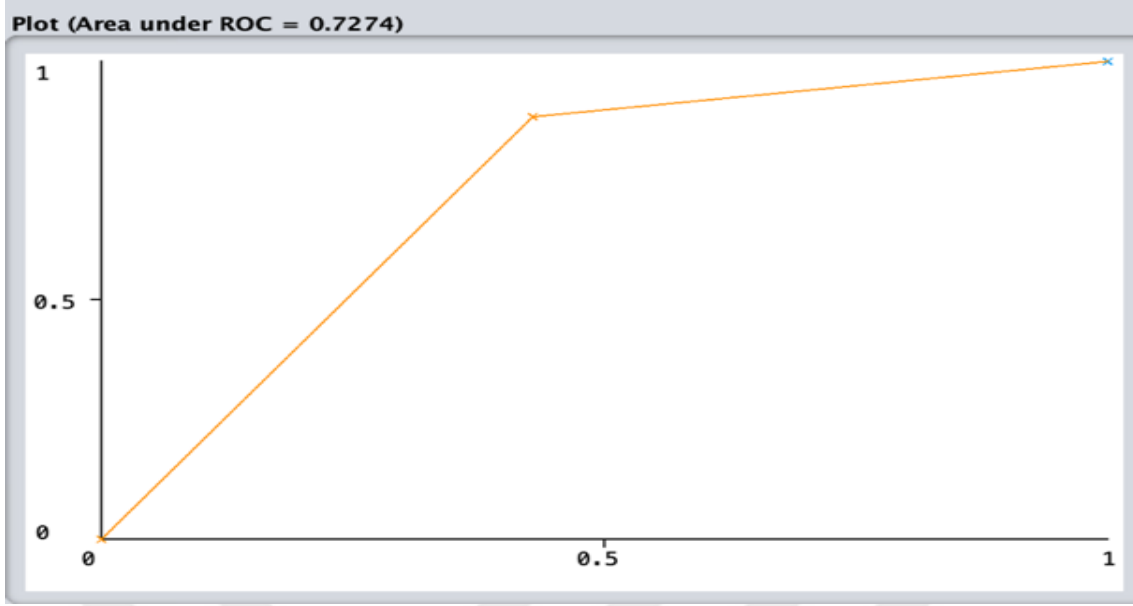
Şekil 4.33'te bad ve god örneklerine ait ROC eğrisinin altında kalan alanın büyük olduğu görülmektedir. Ayrıca ROC değeri 0.7 elde edilmiştir. Modelin örnekleri sınıflandırmada başarılı olduğu söylenebilir.

4.8 Pima Diabets Veri Seti Analiz Sonuçları

Çizelge 4.34 Pima diabets veri seti lineer çekirdek fonksiyonu analiz sonucu

Tested negative		Tested negative		Kappa istatistik değeri	Doğruluk oranı
tp	tn	tp	tn		
442	58	153	115	0.4785	77.474

Çizelge 4.34'te Lineer çekirdek fonksiyonu yardımıyla analiz edilen verilerde doğruluk oranı 77.474 çıkmıştır. Ayrıca kappa istatistik değerinin 0.4785 çıkması modelin yeterince güvenilir ve sınıflar arasındaki uyumun iyi olmadığı söylenebilir.



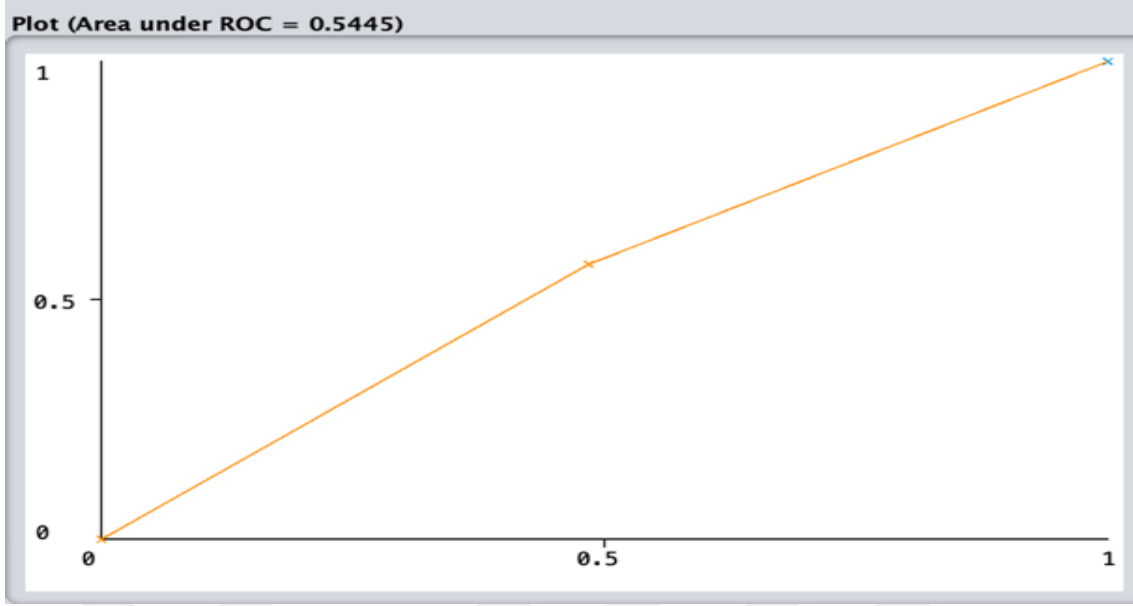
Şekil 4.34 Pima diabets veri seti lineer çekirdek fonksiyonu ile analiz edilen örneklere ait ROC eğrisi

Şekil 4.34'te negative test ve positive test örneklerine ait ROC eğrisinin altında kalan alanın büyük olduğu görülmektedir. Ayrıca ROC değeri 0.7274 elde edilmiştir. Modelin örnekleri sınıflandırmada başarılı olduğu söylenebilir.

Çizelge 4.35 Pima diabets veri seti polinom çekirdek fonksiyonu analiz sonucu

Tested negative		Tested negative		Kappa istatistik değeri	Doğruluk oranı
tp	tn	tp	tn		
287	213	138	130	0.083	55.3385

Çizelge 4.35'te Polinom çekirdek fonksiyonu yardımıyla analiz edilen verilerde doğruluk oranı 55.3385 çıkmıştır. Ayrıca kappa istatistik değerinin 0.083 çıkması modelin güvenilirmez ve sınıflar arasındaki uyumun iyi olmadığı söylenebilir.



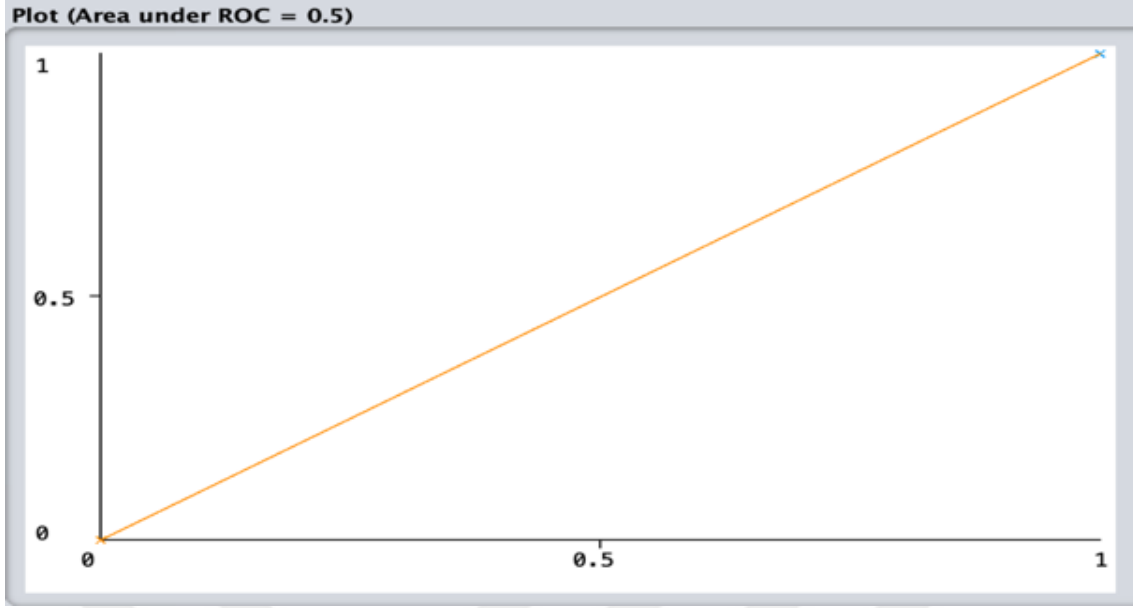
Şekil 4.35 Pima diabetis veri seti Polinom çekirdek fonksiyonu ile analiz edilen örneklere ait ROC eğrisi

Şekil 4.35'te negative test ve positive test örneklerine ait ROC eğrisinin altında kalan alanın küçük olduğu görülmektedir. Ayrıca ROC değeri 0.5445 elde edilmiştir. Modelin örnekleri sınıflandırmada başarısız olduğu söylenebilir.

Çizelge 4.36 Pima diabetis veri seti RBF ve Sigmoid çekirdek fonksiyonu analiz sonucu

Tested negative		Tested negative		Kappa istatistik değeri	Doğruluk oranı
tp	tn	tp	tn		
500	0	0.	268	0.349	65.1042

Çizelge 4.36'da RBF ve Sigmoid çekirdek fonksiyonu yardımıyla analiz edilen verilerde doğruluk oranı 65.1042 çıkmıştır. Ayrıca kappa istatistik değerinin 0.349 çıkması modelin güvenilirmez ve sınıflar arasındaki uyumun iyi olmadığı söylenebilir.



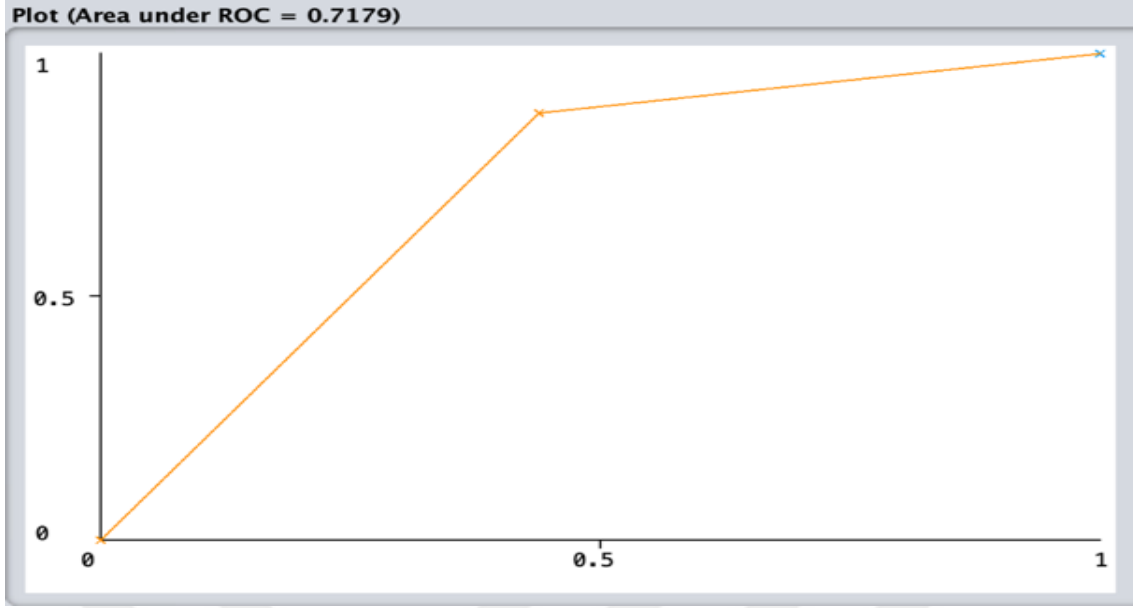
Şekil 4.36 Pima diabets veri seti RBF ve sigmoid çekirdek fonksiyonu ile analiz edilen örneklerle ait ROC eğrisi

Şekil 4.36’da negative test ve positive test örneklerine ait ROC eğrisinin altında kalan alanın küçük olduğu görülmektedir. Ayrıca ROC değeri 0.5 elde edilmiştir. Modelin örnekleri sınıflandırmada başarısız olduğu söylenebilir.

Çizelge 4.37 Pima diabets veri seti Pearson VII işlevsel çekirdek fonksiyonu analiz sonucu

Tested negative		Tested negative		Kappa istatistik değeri	Doğruluk oranı
tp	tn	tp	tn		
438	62	150	118	0.4579	76.5625

Çizelge 4.37’de Pearson VII işlevsel çekirdek fonksiyonu yardımıyla analiz edilen verilerde doğruluk oranı 76.5625 çıkmıştır. Ayrıca kappa istatistik değerinin 0.4579 çıkması modelin güvenilirmez ve sınıflar arasındaki uyumun iyi olmadığı söylenebilir.



Şekil 4.37 Pima diabetis veri seti Pearson VII işlevsel çekirdek fonksiyonu ile analiz edilen örneklere ait ROC eğrisi

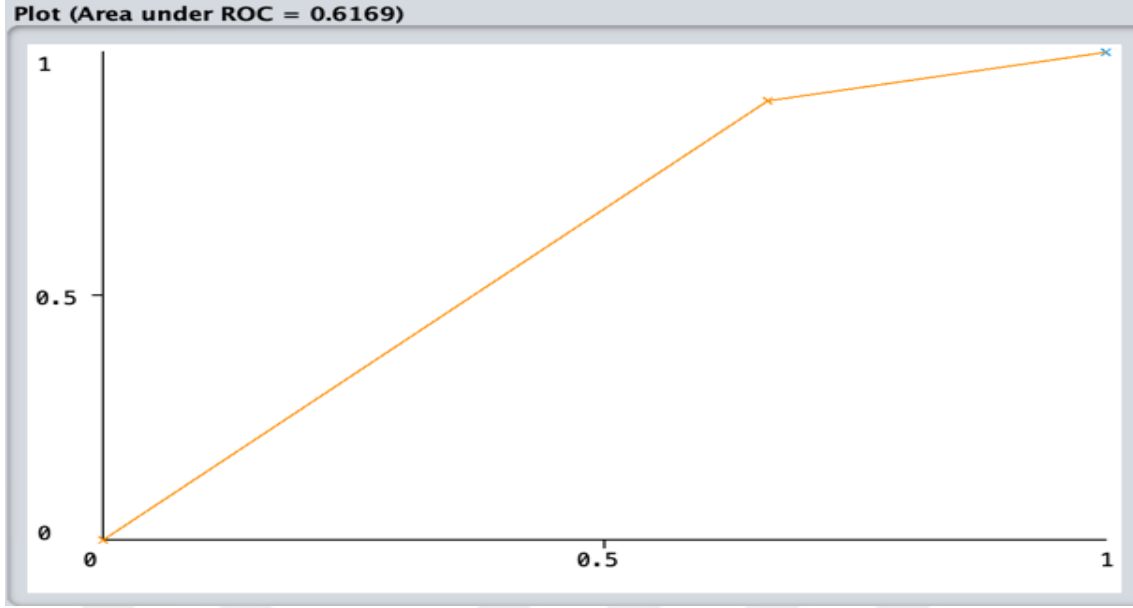
Şekil 4.37’de negative test ve positive test örneklerine ait ROC eğrisinin altında kalan alanın büyük olduğu görülmektedir. Ayrıca ROC değeri 0.7179 elde edilmiştir. Modelin örnekleri sınıflandırmada başarılı olduğu söylenebilir.

4.9 German Credit Veri Seti Analiz Sonuçları

Çizelge 4.38 German credit veri seti Lineer çekirdek fonksiyonu analiz sonucu

good		bad		kappa istatistik değeri	doğruluk oranı
tp	tn	tp	tn		
628	72	101	199	0.266	72.9

Çizelge 4.38’de lineer çekirdek fonksiyonu yardımıyla analiz edilen verilerde doğruluk oranı 72.9 çıkmıştır. Ayrıca kappa istatistik değerinin 0.266 çıkması modelin güvenilirmez ve sınıflar arasındaki uyumun iyi olmadığı söylenebilir.



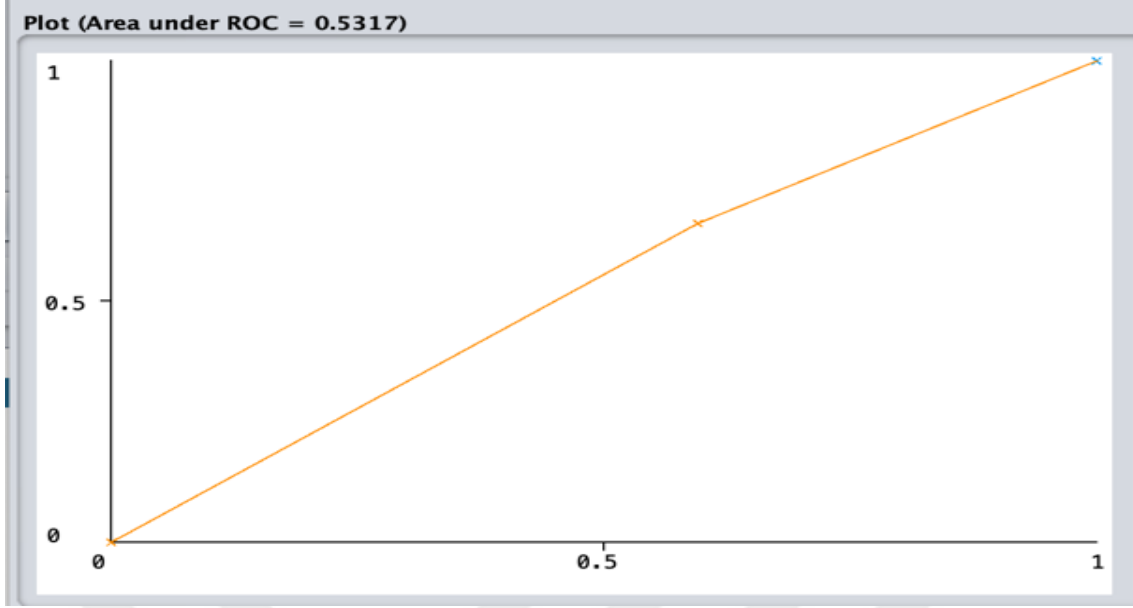
Şekil 4.38 German credit veri seti lineer çekirdek fonksiyonu ile analiz edilen örneklere ait ROC eğrisi

Şekil 4.38’de good ve bad örneklerine ait ROC eğrisinin altında kalan alanın büyük olduğu görülmektedir. Ayrıca ROC değeri 0.6169 elde edilmiştir. Modelin örnekleri sınıflandırmada başarılı olduğu söylenebilir.

Çizelge 4.39 German credit veri seti Polinom çekirdek fonksiyonu analiz sonucu

good		bad		kappa istatistik değeri	doğruluk oranı
tp	tn	tp	tn		
462	238	121	179	0.417	58.3

Çizelge 4.39’da polinom çekirdek fonksiyonu yardımıyla analiz edilen verilerde doğruluk oranı 58.3 çıkmıştır. Ayrıca kappa istatistik değerinin 0.417 çıkması modelin güvenilir ve sınıflar arasındaki uyumun iyi olmadığı söylenebilir.



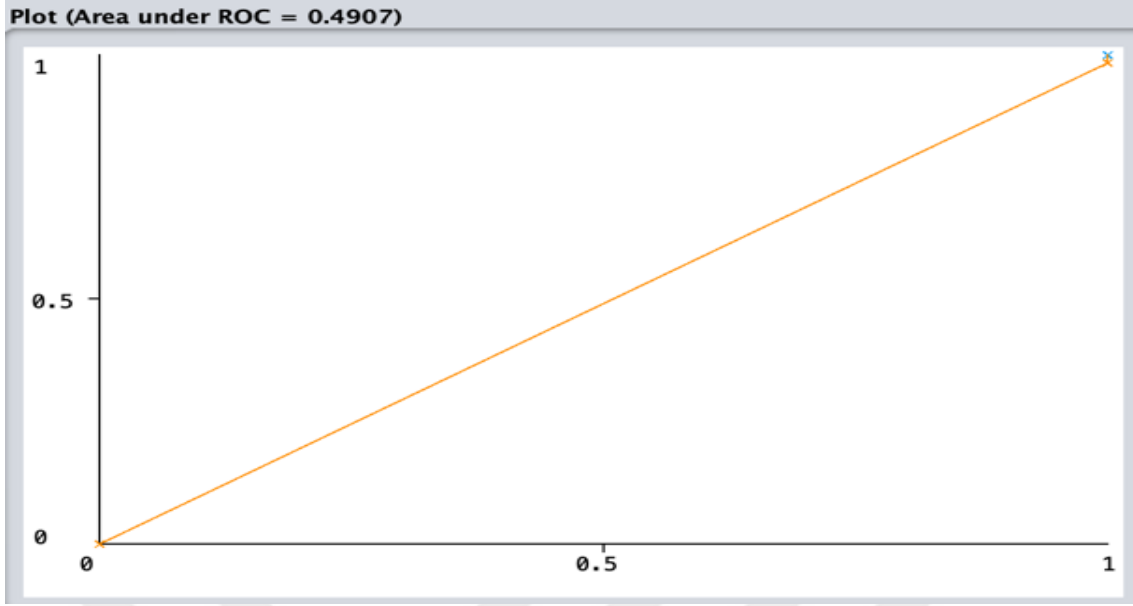
Şekil 4.39 German credit veri seti polinom çekirdek fonksiyonu ile analiz edilen örneklere ait ROC eğrisi

Şekil 4.39’da good ve bad örneklerine ait ROC eğrisinin altında kalan alanın küçük olduğu görülmektedir. Ayrıca ROC değeri 0.5317 elde edilmiştir. Modelin örnekleri sınıflandırmada başarısız olduğu söylenebilir.

Çizelge 4.40 German credit veri seti RBF çekirdek fonksiyonu analiz sonucu

good		bad		kappa istatistik değeri	doğruluk oranı
tp	tn	tp	tn		
687	13	300	0	-0.0256	68.7

Çizelge 4.40’ta RBF çekirdek fonksiyonu yardımıyla analiz edilen verilerde doğruluk oranı 68.7 çıkmıştır. Ayrıca kappa istatistik değerinin -0.0256 çıkması modelin güvenilirmez ve sınıflar arasındaki uyumun iyi olmadığı söylenebilir.



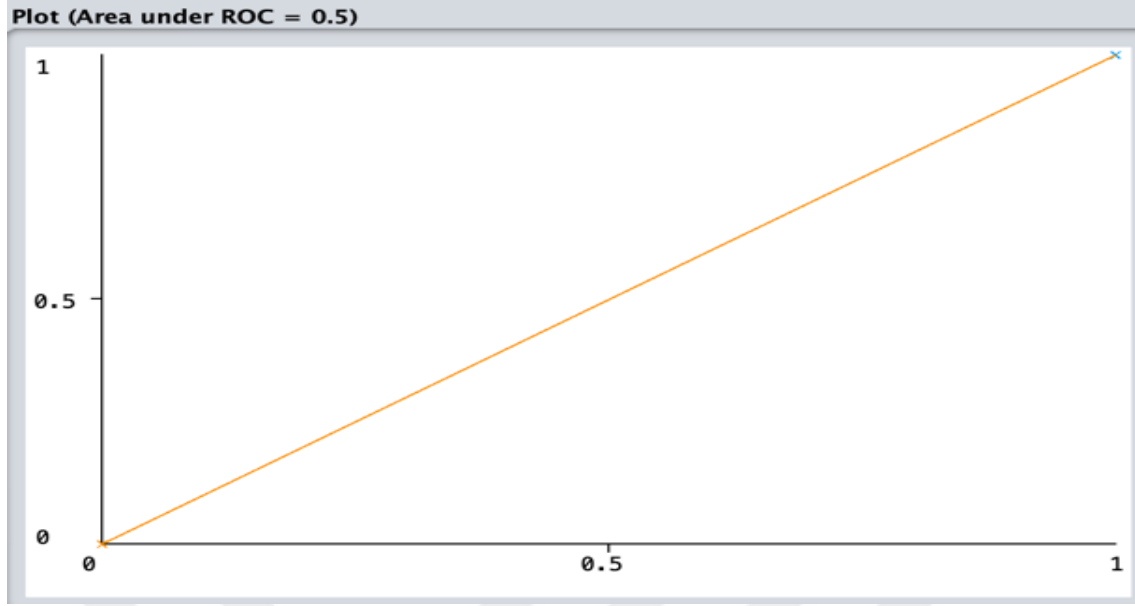
Şekil 4.40 German credit veri seti RBF çekirdek fonksiyonu ile analiz edilen örneklere ait ROC eğrisi

Şekil 4.40'ta good ve bad örneklerine ait ROC eğrisinin altında kalan alanın küçük olduğu görülmektedir. Ayrıca ROC değeri 0.4907 elde edilmiştir. Modelin örnekleri sınıflandırmada başarısız olduğu söylenebilir.

Çizelge 4.41 German credit veri seti Sigmoid çekirdek fonksiyonu analiz sonucu

good		bad		kappa istatistik değeri	doğruluk oranı
tp	tn	tp	tn		
700	0	0	300	0	70

Çizelge 4.41'de sigmoid çekirdek fonksiyonu yardımıyla analiz edilen verilerde doğruluk oranı 70 çıkmıştır. Ayrıca kappa istatistik değerinin 0 çıkması modelin güvenilirmez ve sınıflar arasındaki uyumun iyi olmadığı söylenebilir.



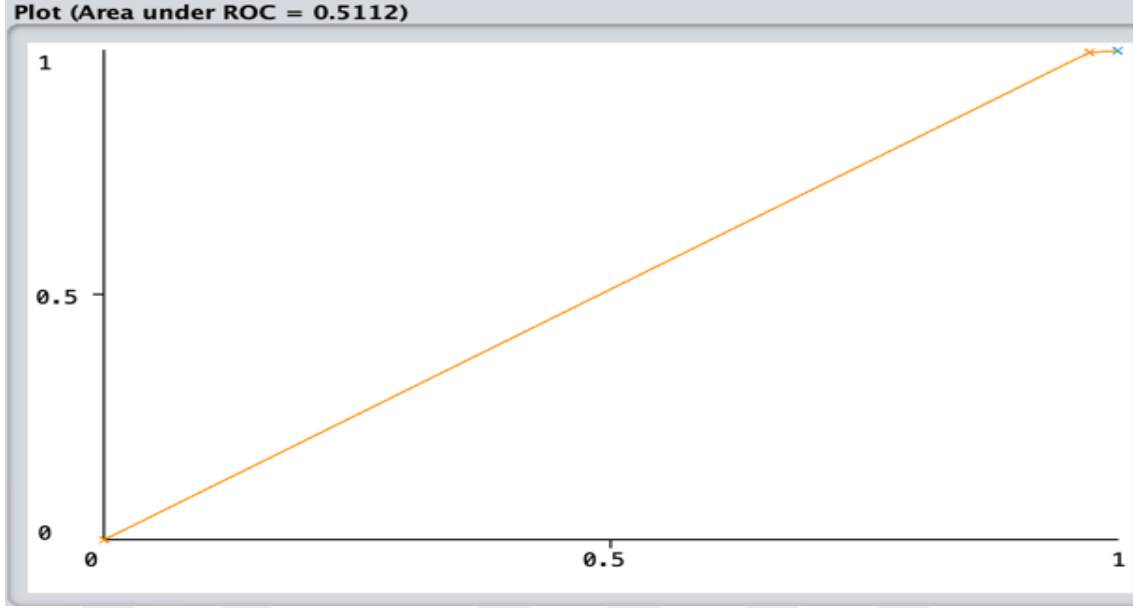
Şekil 4.41 German credit veri seti sigmoid çekirdek fonksiyonu ile analiz edilen örneklere ait ROC eğrisi

Şekil 4.41’de good ve bad örneklerine ait ROC eğrisinin altında kalan alanın küçük olduğu görülmektedir. Ayrıca ROC değeri 0.5 elde edilmiştir. Modelin örnekleri sınıflandırmada başarısız olduğu söylenebilir.

Çizelge 4.42 German credit veri seti Pearson VII işlevsel çekirdek fonksiyonu analiz sonucu

good		bad		kappa istatistik değeri	doğruluk oranı
tp	tn	tp	tn		
697	3	8	292	0.0309	70.5

Çizelge 4.42’de Pearson VII işlevsel çekirdek fonksiyonu yardımıyla analiz edilen verilerde doğruluk oranı 70.5 çıkmıştır. Ayrıca kappa istatistik değerinin 0.0309 çıkması modelin güvenilmez ve sınıflar arasındaki uyumun iyi olmadığı söylenebilir.



Şekil 4.42 German credit veri seti pearson VII işlevsel çekirdek fonksiyonu ile analiz edilen örneklere ait ROC eğrisi

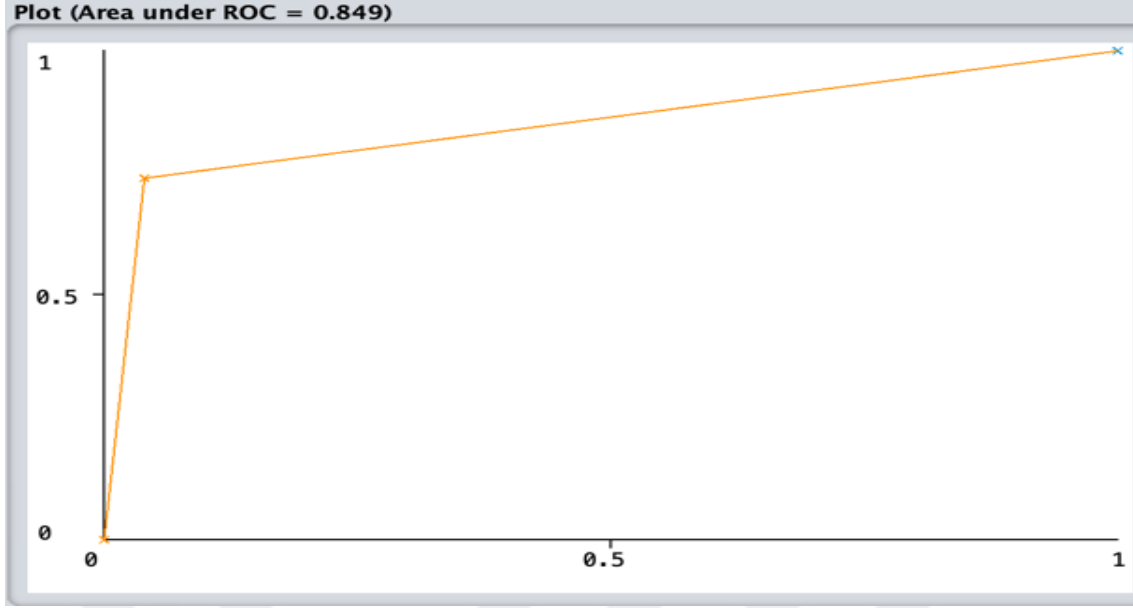
Şekil 4.42’de good ve bad örneklerine ait ROC eğrisinin altında kalan alanın küçük olduğu görülmektedir. Ayrıca ROC değeri 0.5112 elde edilmiştir. Modelin örnekleri sınıflandırmada başarısız olduğu söylenebilir.

4.10 Inosphere Veri Seti Analiz Sonuçları

Çizelge 4.43 Ionosphere veri seti Lineer çekirdek fonksiyonu analiz sonucu

a		b		kappa istatistik değeri	doğruluk oranı
tp	tn	tp	tn		
93	33	216	9	0.7286	88.0342

Çizelge 4.43’te lineer çekirdek fonksiyonu yardımıyla analiz edilen verilerde doğruluk oranı 88.0342 çıkmıştır. Ayrıca kappa istatistik değerinin 0.7286 çıkması modelin güvenilir ve sınıflar arasındaki uyumun iyi olduğu söylenebilir.



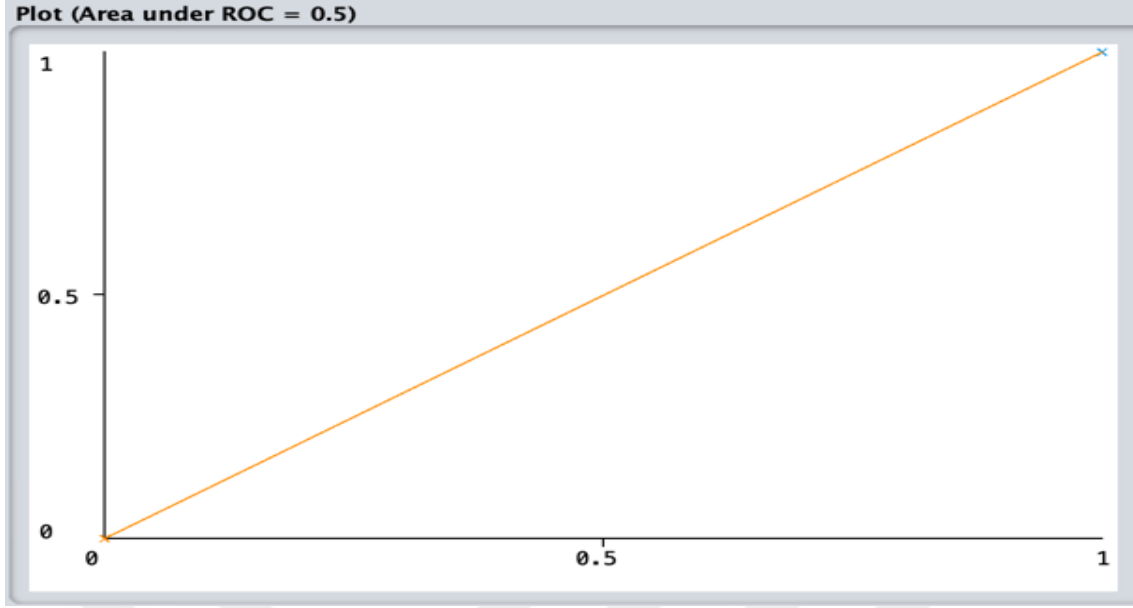
Şekil 4.43 Ionosphere veri seti lineer çekirdek fonksiyonu ile analiz edilen örneklere ait ROC eğrisi

Şekil 4.43'te a ve b örneklerine ait ROC eğrisinin altında kalan alanın büyük olduğu görülmektedir. Ayrıca ROC değeri 0.849 elde edilmiştir. Modelin örnekleri sınıflandırmada başarılı olduğu söylenebilir.

Çizelge 4.44 Ionosphere veri seti Polinom çekirdek fonksiyonu analiz sonucu

a		b		kappa istatistik değeri	doğruluk oranı
tp	tn	tp	tn		
0	126	225	0	0	64.1026

Çizelge 4.44'te polinom çekirdek fonksiyonu yardımıyla analiz edilen verilerde doğruluk oranı 64.1026 çıkmıştır. Ayrıca kappa istatistik değerinin 0 çıkması modelin güvenilirmez ve sınıflar arasındaki uyumun iyi olmadığı söylenebilir.



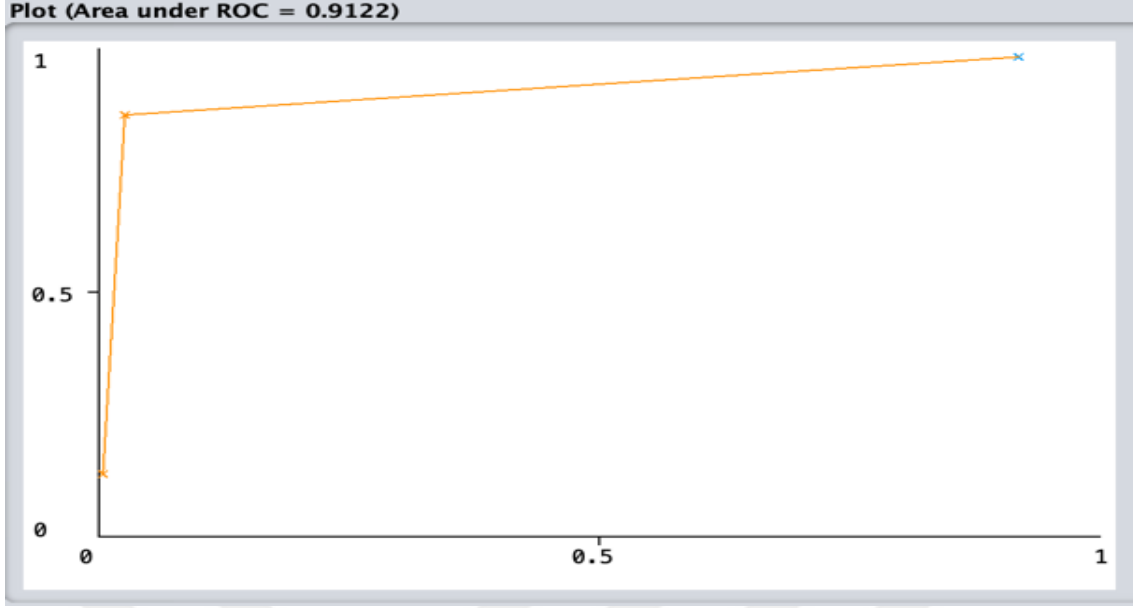
Şekil 4.44 Ionosphere veri seti polinom çekirdek fonksiyonu ile analiz edilen örneklere ait ROC eğrisi

Şekil 4.44'te a ve b örneklerine ait ROC eğrisinin altında kalan alanın küçük olduğu görülmektedir. Ayrıca ROC değeri 0.5 elde edilmiştir. Modelin örnekleri sınıflandırmada başarısız olduğu söylenebilir.

Çizelge 4.45 Ionosphere veri seti RBF çekirdek fonksiyonu analiz sonucu

a		b		kappa istatistik değeri	doğruluk oranı
tp	tn	tp	tn		
105	21	223	2	0.8527	93.4473

Çizelge 4.45'te RBF çekirdek fonksiyonu yardımıyla analiz edilen verilerde doğruluk oranı 93.4473 çıkmıştır. Ayrıca kappa istatistik değerinin 0.8527 çıkması modelin güvenilir ve sınıflar arasındaki uyumun iyi olduğu söylenebilir.



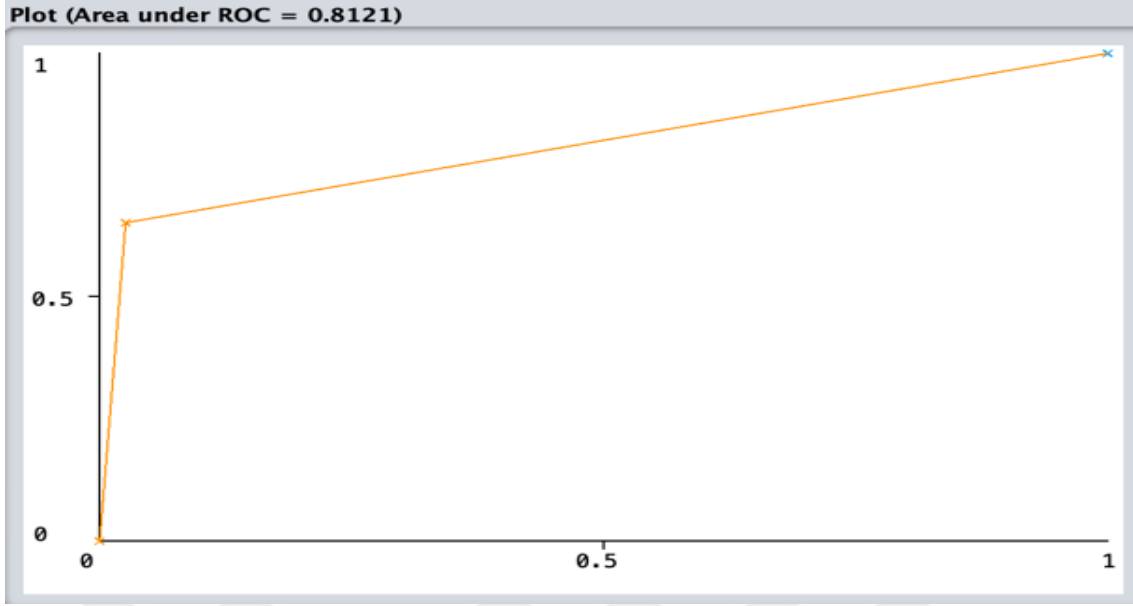
Şekil 4.45 Ionosphere veri seti RBF çekirdek fonksiyonu ile analiz edilen örneklere ait ROC eğrisi

Şekil 4.45'te a ve b örneklerine ait ROC eğrisinin altında kalan alanın büyük olduğu görülmektedir. Ayrıca ROC değeri 0.9122 elde edilmiştir. Modelin örnekleri sınıflandırmada başarılı olduğu söylenebilir.

Çizelge 4.46 Ionosphere veri seti Sigmoid çekirdek fonksiyonu analiz sonucu

a		b		kappa istatistik değeri	doğruluk oranı
tp	tn	tp	tn		
82	44	219	6	0.6685	85.755

Çizelge 4.46'da sigmoid çekirdek fonksiyonu yardımıyla analiz edilen verilerde doğruluk oranı 85.755 çıkmıştır. Ayrıca kappa istatistik değerinin 0.6685 çıkması modelin güvenilir ve sınıflar arasındaki uyumun iyi olduğu söylenebilir.



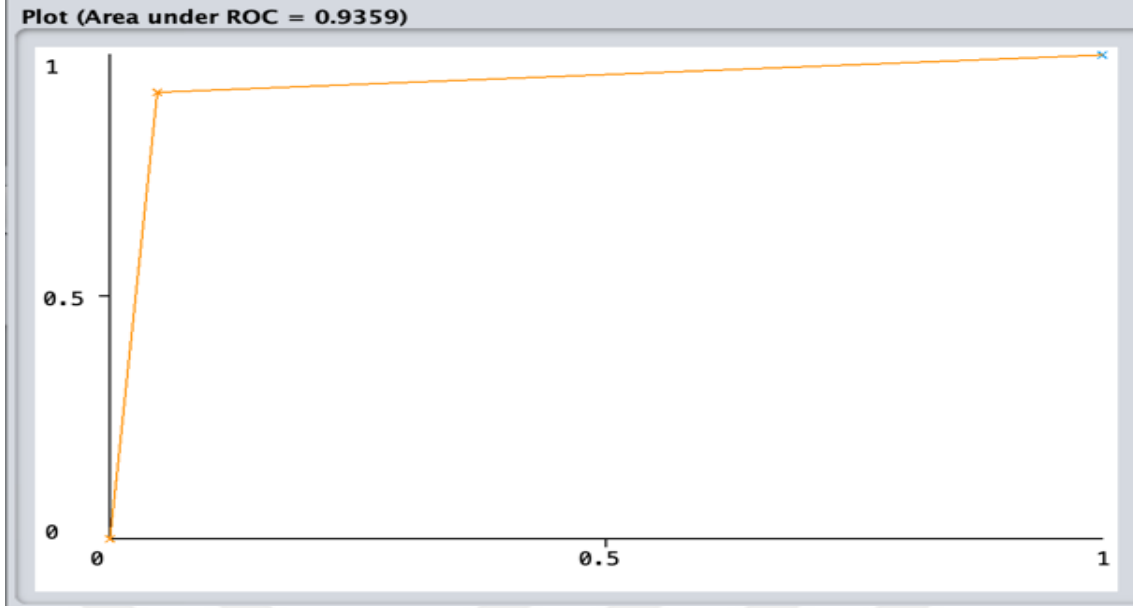
Şekil 4.46 Ionosphere veri seti sigmoid çekirdek fonksiyonu ile analiz edilen örneklere ait ROC eğrisi

Şekil 4.46’da a ve b örneklerine ait ROC eğrisinin altında kalan alanın büyük olduğu görülmektedir. Ayrıca ROC değeri 0.8121 elde edilmiştir. Modelin örnekleri sınıflandırmada başarılı olduğu söylenebilir.

Çizelge 4.47 Ionosphere veri seti Pearson VII işlevsel çekirdek fonksiyonu analiz sonucu

a		b		kappa istatistik değeri	doğruluk oranı
tp	tn	tp	tn		
116	10	214	11	0.8702	94.0171

Çizelge 4.47’de Pearson VII işlevsel çekirdek fonksiyonu yardımıyla analiz edilen verilerde doğruluk oranı 94.017 çıkmıştır. Ayrıca kappa istatistik değerinin 0.8702 çıkması modelin güvenilir ve sınıflar arasındaki uyumun iyi olduğu söylenebilir.



Şekil 4.47 Ionosphere veri seti pearson VII işlevsel çekirdek fonksiyonu ile analiz edilen örneklere ait ROC eğrisi

Şekil 4.47’de a ve b örneklerine ait ROC eğrisinin altında kalan alanın büyük olduğu görülmektedir. Ayrıca ROC değeri 0.9359 elde edilmiştir. Modelin örnekleri sınıflandırmada başarılı olduğu söylenebilir.

Çizelge 4.48 Göğüs kanseri veri setinin çekirdek fonksiyonları ile elde edilen analiz sonuçları

Çekirdek Fonksiyonları	Doğruluk oranı	Kappa istatistiği	ROC (AUC) eğrisi değeri	C parametresi
Lineer çekirdek	%97.891	0.973	0.9727	1
Polinom çekirdek	%97.7253	0.9507	0.9713	10000
Radial Basis çekirdek	%98.06	0.9584	0.976	200
Sigmoid çekirdek	%97.891	0.9545	0.9723	24
Pearson VII İşlevsel	%97.5395	0.9471	0.9705	1

Çizelge 4.49 Otizm taraması yetişkin veri setinin çekirdek fonksiyonları ile elde edilen analiz sonuçları

Çekirdek Fonksiyonları	Doğruluk oranı	Kappa istatistiği	ROC (AUC) eğrisi değeri	C parametresi
Lineer çekirdek	%100	1	1	1
Polinom çekirdek	%99.14	0.9784	0.9925	1
Radial Basis çekirdek	%92.898	0.9818	0.9868	1
Sigmoid çekirdek	%73.15	0	0.5	1
Pearson VII İşlevsel	%94.031	0.8384	0.8939	1

Çizelge 4.50 Raisin veri setinin çekirdek fonksiyonları ile elde edilen analiz sonuçları

Çekirdek Fonksiyonları	Doğruluk oranı	Kappa istatistiği	ROC (AUC) eğrisi değeri	C parametresi
Lineer çekirdek	%85.22	0.7044	0.8522	1
Polinom çekirdek	%75.44	0.5089	0.7544	1
Radial Basis çekirdek	%50.33	0.67	0.5033	1
Sigmoid çekirdek	%75	0	0.5	1
Pearson VII İşlevsel	%86.6667	0.7333	0.8667	1

Çizelge 4.51 Voice veri setinin çekirdek fonksiyonları ile elde edilen analiz sonuçları

Çekirdek Fonksiyonları	Doğruluk oranı	Kappa istatistiği	ROC (AUC) eğrisi değeri	C parametresi
Lineer çekirdek	%92.329	0.8466	0.9333	1
Polinom çekirdek	%57.89	0.15	0.5789	1
Radial Basis çekirdek	%73.67	0.4735	0.7367	1
Sigmoid çekirdek	%43.46	0.1307	0.4347	1
Pearson VII işlevsel.	%62.7525.	0.2551	0.6275	1

Çizelge 4.52 Weather veri setinin çekirdek fonksiyonları ile elde edilen analiz sonuçları

Çekirdek Fonksiyonları	Doğruluk oranı	Kappa istatistiği	ROC (AUC) eğrisi değeri	C parametresi
Lineer çekirdek	%57.142	0.0667	0.5333	1
Polinom çekirdek	%71.4286	0.3778	0.6889	1
Radial Basis çekirdek	%54.2857	0	0.5	1
Sigmoid çekirdek	%43.46	0.1307	0.4347	1
Pearson VII işlevsel.	%57.1429	-0.1351	0.4444.	1

Çizelge 4.53 Breast cancer veri setinin çekirdek fonksiyonları ile elde edilen analiz sonuçları

Çekirdek Fonksiyonları	Doğruluk oranı	Kappa istatistiği	ROC (AUC) eğrisi değeri	C parametresi
Lineer çekirdek	%69.2308	0.19	0.5876	1
Polinom çekirdek	%70.2797	0	0.5	1
Radial Basis çekirdek	%70.6294	0.257	0.5093	1
Sigmoid çekirdek	%70.2797	0	0.5	1
Pearson VII işlevsel.	%70.2797	0	0.5	1

Çizelge 4.54 Labor neg veri setinin çekirdek fonksiyonları ile elde edilen analiz sonuçları

Çekirdek Fonksiyonları	Doğruluk oranı	Kappa istatistiği	ROC (AUC) eğrisi değeri	C parametresi
Lineer çekirdek	%92.9825	0.8423	0.9115	1
Polinom çekirdek	%91.2281	0.8052	0.898	1
Radial Basis çekirdek	%92.9825	0.8423	0.9115	1
Sigmoid çekirdek	%64.9123	0.3507	0.5	1
Pearson VII işlevsel.	%78.9474	0.4639	0.7	1

Çizelge 4.55 Pima diabets veri setinin çekirdek fonksiyonları ile elde edilen analiz sonuçları

Çekirdek Fonksiyonları	Doğruluk oranı	Kappa istatistiği	ROC (AUC) eğrisi değeri	C parametresi
Lineer çekirdek	%77.47	0.4785	0.7274	1
Polinom çekirdek	%55.3385	0.083	0.5445	1
Radial Basis çekirdek	%65.1042	0.349	0.5	1
Sigmoid çekirdek	%65.1042	0.349	0.5	1
Pearson VII işlevsel.	%76.5625	0.4579	0.7179	1

Çizelge 4.56 German credit veri setinin çekirdek fonksiyonları ile elde edilen analiz sonuçları

Çekirdek Fonksiyonları	Doğruluk oranı	Kappa istatistiği	ROC (AUC) eğrisi değeri	C parametresi
Lineer çekirdek	%72.9	0.266	0.6169	1
Polinom çekirdek	%58.3	0.417	0.5317	1
Radial Basis çekirdek	%68.7	-0.02569	0.4907	1
Sigmoid çekirdek	% 70	0	0.5	1
Pearson VII işlevsel.	%70.5	0.0309	0.5112	1

Çizelge 4.57 Ionosphere veri setinin çekirdek fonksiyonları ile elde edilen analiz sonuçları

Çekirdek Fonksiyonları	Doğruluk oranı	Kappa istatistiği	ROC (AUC) eğrisi değeri	C parametresi
Lineer çekirdek	%88.0342	0.7286	0.849	1
Polinom çekirdek	%62.1026	0	0.5	1
Radial Basis çekirdek	%93.4473	0.8527	0.9122	1
Sigmoid çekirdek	%85.755	0.6685	0.8121	1
Pearson VII işlevsel.	%94.071	0.8702	0.9359	1

Çizelge 4.58 Veri setlerine ait özellikler ve en iyi performans gösteren çekirdek fonksiyonları

Veri setleri	Performans sonucu en iyi kernel	veri çeşidi
Göğüs kanseri veri seti	RBF	sürekli
Otizm taraması yetişkin veri seti	Lineer	kesikli, sürekli, string
Raisin veri seti	Pearson VII işlevsel	sürekli
Voice veri seti	Lineer	sürekli
Weather veri seti	Polinom	kesikli, string
Breast cancer veri seti	RBF	kesikli, string
Labor neg veri seti	Lineer ve RBF	sürekli, kesikli, string
Pima diabets veri seti	Pearson VII işlevsel	sürekli, kesikli
German Credit veri seti.	Pearson VII işlevsel	kesikli, string
Ionosphere veri seti	Pearson VII işlevsel	sürekli



5. TARTIŞMA SONUÇ

SVM algoritması ile 10 adet veri seti analiz edilmiştir. Bu analiz sonucunda en yüksek sınıflandırma başarısı; 3 adet kesikli veri setinde RBF, lineer ve pearson VII işlevsel çekirdek fonksiyonları, hem sürekli hem de kesikli olan 1 adet veri setinde iyi sınıflandırma başarısı pearson VII işlevsel çekirdek fonksiyonu, kesikli, sürekli ve string verilerden oluşan 2 adet veri seti için en iyi sınıflandırma başarısı lineer ve RBF çekirdek fonksiyonu, kesikli ve string verilerden oluşan 2 adet veri setinde en yüksek sınıflandırma başarısı ise polinom, pearson VII işlevsel ve RBF çekirdek fonksiyonları ile elde edilmiştir. Tüm bu sonuçlar incelendiğinde lineer ve pearson VII işlevsel çekirdek fonksiyonunun sürekli verilerin bulunduğu veri setlerinde daha başarılı sonuçlar elde ettiği gözlenmiştir. Ayrıca kesikli, sürekli ve string verilerin bulunduğu veri setlerinde lineer çekirdeğin yine başarılı sonuçlar elde ettiği gözlenmiştir. RBF çekirdek fonksiyonunun ise kesikli veri setlerinde daha başarılı olduğu gözlenmiştir. Pearson VII işlevsel çekirdek fonksiyonunun sadece sürekli verilerin bulunduğu veri setinde başarılı sonuçlar elde ettiği gözlenmiştir. Lineer çekirdek ve pearson VII işlevsel çekirdek fonksiyonunun genel olarak daha başarılı sonuçlar elde ettiği gözlenmiştir.

Göğüs kanseri veri seti için kullanılan çekirdek fonksiyonları için, RBF için C parametresinin en ideal değerinin ise ikiyüz (200), polinom çekirdek fonksiyonu için en iyi C değerinin on bin (10000), sigmoid çekirdek fonksiyonu için en iyi C değeri yirmi dört (24) olduğu deneme yanılma yöntemi ile tespit edilmiştir. Diğer çekirdek fonksiyonları ve göğüs kanseri veri dışındaki veri setleri için C değeri bir (1) seçilmiştir. Ayrıca kullanılan çekirdek fonksiyonları ile kurulan modelin iyi olup olmadığı sınıflar arasında ki uyumun nasıl olduğu gözlenmiştir. Kullanılan çekirdek fonksiyonlarında kappa istatistiği ve ROC değerleri incelendiğinde hem yüksek hem de düşük oranlar elde edildiği tespit edilmiştir. Çekirdek fonksiyonları parametreleri sabit tutulmuştur. Literatürde benzer çalışmalar incelenmiş ve bu çalışmalarda; Ayhan ve Erdoğan (2014)'nın 12 veri seti üzerinde SVM algoritması ile sınıflandırma yapmışlardır. Genel olarak en iyi sınıflandırma başarısını Radial Basis çekirdek fonksiyonu ile farklı veri tiplerinde elde etmişlerdir. Tarihi ve ark. (2020) bankaların başarısızlıkları üzerine yaptıkları çalışmada en iyi sınıflandırma başarısını SVM algoritması ve Radial Basis çekirdek fonksiyonu ile elde etmişlerdir. (Aydınadağ Bayrak ve ark., 2021) göğüs kanseri

verisi üzerine yaptıkları çalışmada SVM algoritmasını kullanmışlardır. C parametresinin değerini 1 ve çekirdek fonksiyonu parametrelerini sabit tutmuşlardır. En iyi sınıflandırma başarısını Radial Basis ve Gamma çekirdek fonksiyonunu ile elde etmişlerdir.

Destek vektör makineleri ve çekirdek fonksiyonları kullanılarak oluşturulan sınıflandırma modelinde, sonradan modele dahil olacak verileri belirli bir hata oranında sınıflandırabilir. Ayrıca veri setlerinin özelliklerine bağlı olarak hangi çekirdek fonksiyonunu kullanacaklarına karar vermede destek olunabilir. Benzer çalışmalarda bulunacak kişiler hem kullanacakları çekirdek fonksiyonları hem de bu model sayesinde ön bilgiye sahip olunabilir. Bu sayede verilerin sınıflandırılmasında minimum hata ve yüksek başarı oranı elde edilebilir.

KAYNAKLAR

- Abbasi, A., Zhang, Z., Zimbra, D., Chen, H., Nunamaker, J. F. (2010). Detecting fake websites the contribution of statistical learning theory. *MIS Quarterly: Management Information Systems*, 34(3), 435-461.
- Abdi, H., Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews Computational Statistics*, 2(4), 433-459.
- Aydındağ Bayrak, E., Kırıcı, P., Ensari, T., Seven, E., Mustafa, D. (2022). Göğüs Kanseri Verileri Üzerinde Makine Öğrenmesi Yöntemlerinin Uygulanması. *Journal of Intelligent Systems Theory and Applications*, 5(1), 35-41.
- Aydın, N., Atasoy, D. T. (2018). Destek Vektör Makineleri Kullanarak Yüz Tanıma Uygulaması Geliştirilmesi. *Comprehensive Physiology*, 6(19), 387-387.
- Ayhan, S., Erdoğan, Ş. (2014a). Destek Vektör Makineleriyle Sınıflandırma Problemlerinin Çözümü İçin Çekirdek Fonksiyonu Seçimi. *Eskişehir Osmangazi Üniversitesi İİBF Dergisi*, 9, 175-198.
- Ayhan, S., Erdoğan, Ş. (2014b). Destek Vektör Makineleriyle Sınıflandırma Problemlerinin Çözümü İçin Çekirdek Fonksiyonu Seçimi. *Eskişehir Osmangazi Üniversitesi İİBF Dergisi*, 9(1), 175-198.
- Babaei, M., Moeini, R., Ehsanzadeh, E. (2019). Artificial Neural Network and Support Vector Machine Models for Inflow Prediction of Dam Reservoir. *Department of Civil Engineering*, 33, 2203-2218.
- Ben-Hur, A., Ong, C. S., Sonnenburg, S., Schölkopf, B., Rätsch, G. (2008). Support vector machines and kernels for computational biology. *PLoS Computational Biology*, 4(10).
- Bilgen, B., Özge Doğa, N. (2017). Puanlayıcılar arası güvenilirlik belirleme tekniklerinin karşılaştırılması. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, 8(1), 63-63.
- Bilişik, M. T. (2011). Destek vektör makinesi, çoklu regresyon ve doğrusal olmayan programlama ile perakendecilik sektöründe gelir yönetimi için dinamik fiyatlandırma. **6. Üretim Araştırmaları Sempozyumu**, İstanbul, Türkiye.
- Bousquet, O., Boucheron, S., Lugosi, G. (2004). Introduction to Statistical Learning Theory. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 3176, 169-207.
- Bredensteiner, E. J., Bennett, K. P. (1999). Multicategory Classification by Support Vector Machines. *Computational Optimization and Applications*, 12(3), 53-79.
- Brunato, M., Battiti, R. (2005). Statistical learning theory for location fingerprinting in wireless LANs. *Computer Networks*, 47(6), 825-845.
- Budak, H. (2018). Özellik seçim yöntemleri ve yeni bir yaklaşım. *Fen Bilimleri Enstitüsü Dergisi*, 22, 21-31.
- Burges, C. J. C. (1998a). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2), 121-167.
- Burges, C. J. C. (1998b). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2, 121-167.
- Cervantes, J., Garcia Lamont, F., Rodríguez Mazahua, L., Lopez, A. (2019). A comprehensive survey on support vector machine classification applications, challenges and trends. *Neurocomputing*, 408, 189-215.
- Cherkassky, V., Mulier, F. (2007). **Learning from data concepts theory and methods**. John Wiley&Sons: New York, USA.

- Cho, B. H., Yu, H., Lee, J., Chee, Y. J., Kim, I. Y., Kim, S. I. (2008). Nonlinear support vector machine visualization for risk factor analysis using nomograms and localized radial basis function kernels. *IEEE Transactions on Information Technology in Biomedicine*, 12(2), 247-256.
- Chrihtianini, N., Shawe taylor, J. (2000). **An introduction to support vector machine and other kernel-based learning methods**. (1. ed. Vol. 1, pp. 20-60) Cambridge, UK..
- Çiğdem, R. (2016). Roc eğrisi yöntemi ile kesme puanının belirlenmesi. *Academia*, 50, 553-556.
- Cortes, C., Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297.
- Demirel, S., Meral, G., Saraçlı, S. (2020). Destek vektör makineleri ve Türkiye’deki enerji santrallerinde doğal gaz tüketimi üzerine bir uygulama. *Süleyman Demirel University Journal of Natural and Applied Sciences*, 24(2), 411-418.
- Ecemiş, A., Dokuz. A. Ş., Çelik, M. (2019). Çeşitli Veri Kümeleri Üzerinde Veri Madenciliği Algoritmalarının Performansının Değerlendirilmesi, *Uluslararası Türk Dünyası Bilim ve Mühendislik Kongresi*, 5, 2-11.
- Erdal, H. (2015). Makine öğrenmesi yöntemlerinin inşaat sektörüne katkısı: basınç dayanımı tahminlemesi. *Prediction of Compressive Strength*. 21(3), 109-114.
- Evgeniou, T., Pontil, M., Poggio, T. (2000). Statistical learning theory a primer. *International Journal of Computer Vision*, 38(1), 9-13.
- Flake, G. W., Lawrence, S. (2002). Efficient SVM regression training with SMO. *Machine Learning*, 46, 271-290.
- Fletcher, T. (2008). **Support vector machines explained**. İngiltere, UK
- Foody, G. M., Mathur, A. (2004). A relative evaluation of multiclass image classification by support vector machines. *IEEE Transactions on Geoscience and Remote Sensing*, 42(6), 1335-1343.
- Franc, V., Hlaváč, V. (2002). Multi-class support vector machine. *Proceedings - International Conference on Pattern Recognition*, 2, 236–239.
- Gartner, T. (2008). Machine Perceptron Artificial intelligence, **Kernels for structured data** (1. ed., Vol 72, pp. 56-62). Danvers, USA.
- Gärtner, T. (2003). A survey of kernels for structured data. *ACM SIGKDD Explorations Newsletter*, 5(1), 49-58.
- Gökalg, Ö. M. (2022). **Makine Öğrenmesi**. Ankara, Türkiye.
- Güran, A., Uysal, M., Doğrusöz, Ö. (2014). Destek vektör makineleri parametre optimizasyonunun duygu analizi üzerindeki etkisi. *Mühendislik Bilimleri Dergisi*, 16, 48-87.
- Gürsoy, M. I., Subaşı, A. (2008). DVM ile EEG işaretlerinin sınıflandırılmasında TBA, BBA ve DAA’nin performansının karşılaştırılması. *IEEE 16th Signal Processing, Communication and Applications*, 1(3), 211-219.
- Hofmann, T., Schölkopf, B., Smola, A. J. (2008). Kernel methods in machine learning. *Annals of Statistics*, 36(3), 1171–1220.
- Hossin, M., Sulaiman, M. (2015). A Review on evaluation metrics for data. *international Journal of Data Mining & Knowledge Management Process (IJDMP)*, 5(2), 1–11.
- Hsu, C. W., Lin, C. J. 2002. A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, 13(2), 415–425.
- Huan Liu, L. Y. (2005). Toward Integrating Feature Selection Algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering*, 17(4), 491-502.

- Jakkula, V. (2006). *Tutorial on support vector machine (SVM)*. Washington, USA.
- Karakoyun, M., Hacibeyoglu, M. (2018). Biyomedikal veri kümeleri ile makine öğrenmesi sınıflandırma algoritmalarının istatistiksel olarak karşılaştırılması. *DEÜ Mühendislik Bilimler Dergisi*, 16(48), 30-41.
- Karatzoglou, A., Meyer, D., Hornik, K. (2006). Support vector algorithm in R. *Journal of Statistical Software*, 15(9), 1-28.
- Karatzoglou, A., Smola, A., Hornik, K., Zeileis, A. (2004). Kernlab an S4 package for kernel methods in R. *JSS Journal of Statistical Software*, 11, 15-9.
- Kasnavi, S. A., Aminafshar, M., Shariati, M. M., Emam Jomeh Kashan, N., Honarvar, M. (2018). The effect of kernel selection on genome wide prediction of discrete traits by Support Vector Machine. *Gene Reports*, 11, 279-282.
- Kavzoğlu, T., Çölkesen, İ. (2010). Destek Vektör Makineleri ile Uydu Görüntülerinin Sınıflandırılmasında Kernel Fonksiyonlarının Etkilerinin İncelenmesi. *Harita Dergisi*, 144, 73-82.
- Kaynar, O., Yildiz, M., Görmez, Y., Albayrak, A. (2016). Makine öğrenmesi yöntemleri ile duygu analizi. *İşleme Sempozyumu*, 17-18 Eylül 2016, Sivas. 234-241.
- Kecman, V. (2001). *Learning and soft computing: support vector machines, neural networks, and fuzzy logic models*. MIT: London, UK.
- Kim, H. C., Pang, S., Je, H. M., Kim, D., Bang, S. Y. (2003). Constructing support vector machine ensemble. *Pattern Recognition*, 36, 2757-2767.
- Korkmaz, G. (2020). Yapay zekâ yöntemleriyle sınıflandırma ve finans sektöründe kurumsal müşterilere dönük bir uygulama. *Akademik Yaklaşımlar Dergisi*, 11(2), 0-2.
- Kung, S. Y. (2014). *Kernel methods and machine learning. In kernel methods and machine learning*. Cambridge University: New York, USA.
- Lampert, C. H. (2009). Kernel methods in computer vision. *Foundations and Trends in Computer Graphics and Vision*, 4(3), 193-285.
- Li, C. H., Lin, C. T., Kuo, B. C., Chu, H. S. (2010). An automatic method for selecting the parameter of the RBF kernel function to support vector machines. *International Geoscience and Remote Sensing Symposium*, 836-839 Taiwan, Taiwan
- Lin, H. T., Lin, C.J. (2003). *A Study on sigmoid kernels for SVM and the training of non-PSD kernels by SMO-type methods*, Taichung, Taiwan.
- Liu, M., Wang, M., Wang, J., Li, D. (2013). Comparison of random forest, support vector machine and back propagation neural network for electronic tongue data classification: Application to the recognition of orange beverage and Chinese vinegar. *Sensors and Actuators, Chemical*, 177, 970-980.
- Liu, Y., Zheng, Y. F. (2005). One-against-all multi-class SVM classification using reliability measures. *Proceedings of the International Joint Conference on Neural Networks*, 2, 849-854.
- Luxburg, U. von, Schölkopf, B. (2011). Statistical Learning Theory: Models, Concepts, and Results. *Handbook of the History of Logic*, 10, 651-706.
- Madzarov, G., Gjorgjevikj, D., Chorbev, I. (2008). A Multi-class SVM classifier utilizing binary decision tree. *DPLP*, 33(2), 225-233.
- Martin, E., Kaski, S., Zheng, F., Webb, G. I., Zhu, X., Muslea, I., Ting, K. M., Vlachos, M., Miikkulainen, R., Fern, A., Osborne, M., Raedt, L. De, Kersting, K., Zeugmann, T., Zhang, X., Bain, M., Czumaj, A., Sohler, C., Sammut, C., Kersting, K. (2011). Structural risk minimization. *Encyclopedia of Machine Learning*, 5(11), 929-930.

- Metlek, S., Kayaalp, K. (2020). **Destek vektör makineleri**. İksad Yayınevi: Ankara, Türkiye.
- Min, J. H., Lee, Y. C. (2005). Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. *Expert Systems with Applications*, 28(4), 603–614.
- Naganna, S. R., Deka, P. C. (2014). Support vector machine applications in the field of hydrology: A review. *Applied Soft Computing*, 15, 1-15.
- Nanda, M. A., Seminar, K. B., Nandika, D., Maddu, A. (2018). A comparison study of kernel functions in the support vector machine and its application for termite detection. *Information (Switzerland)*, 9(1): 2-14.
- Nitze, I., Schulthess, U. And Asche, H. (2012). Comparison of machine learning algorithms random forest, artificial neural network and support vector machine to maximum likelihood for supervised crop type classification. **Proceedings of the 4th GEOBIA**, Janeiro, Brazil.
- Oğuzlar, A. (2004). Veri ön İŞLEME. *Erciyes Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, 21, 67–76.
- Osowski, S., Siwekand, K., Markiewicz, T. (2004). MLP and SVM Networks – a Comparative Study. **Proceedings of the 6th Nordic Signal Processing Symposium** . Norsig, Norway.
- Özcan, T. (2020). Yığınlanmış özdevinimli kodlayıcılar ile göğüs kanserinin sınıflandırılması ve klasik makine öğrenme metotları ile performans karşılaştırması. *Erciyes University Journal of Institute of Science and Technology*, 36(2), 20-20.
- Özdet, B., İçer, S. (2022). Akciğer bilgisayarlı tomografi görüntülerinde görüntü işleme uygulamaları ile tümörlerinin tespit edilmesi. *Uludağ University Journal of the Faculty of Engineering*, 27(1), 135-150.
- Pérez-Cruz, F., Bousquet, O. (2004). Kernel methods and their potential use in signal processing. *IEEE Signal Processing Magazine*, 21(3), 57-65.
- Pisner, D. A., Schnyer, D. M. (2019). **Support vector machine**. University of Texas. Austin. Texas.
- Platt, J. C., Cristianini, N., Shawe-Taylor, J. (2000). Large margin DAGs for multiclass classification. *Advances in Neural Information Processing Systems*, 547–553.
- Poggio, T., Evgeniou, T., Pontil, M., Verri, A. (2002). Regularization and statistical learning theory for data analysis. *Computational Statistics and Data Analysis*, 38(4), 421-432.
- Pradhan, A. (2012). *Support Vector Machine-A Survey*. *International Journal of Emerging Technology and Advanced Engineering*, 2(8), 82-85.
- Qian, H., Mao, Y., Xiang, W., Wang, Z. 2010. Recognition of human activities using SVM multi-class classifier. *Pattern Recognition Letters*, 31(2), 100-111.
- Richardson, M. (2009). **Principal component analysis**. London, UK.
- Saeys, Y., Inza, I., Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19), 2507–2517.
- Schölkopf, B., Alexander J. Smola. (2002). Learning with kernels. *In Australian Veterinary Journal*, 62, 1.
- Seda Tolun. (2008). *Destek vektör makineleri: banka başarısızlığının tahmini üzerine bir uygulama*, Doktora tezi. İstanbul Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul.
- Selim, K., (2013). ROC analysis in clinical decision making. *Journal of Mood Disorders*, 3(3), 135.

- Sevli, O. (2019). Göğüs kanseri teşhisinde farklı makine öğrenmesi tekniklerinin performans karşılaştırması. *European Journal of Science and Technology*, 16, 176-185.
- Şimşek, O., Gümüş, V., Özlük, A. (2021). Estimation of discharge coefficient of the trapezoidal broad crested weir flow using support vector machines. *Journal of Science*, 9(3), 533-547.
- Smola, A. J., Schölkopf, B., Müller, K. R. (1998). The connection between regularization operators and support vector kernels. *Neural Network*, 11, 637-649.
- Song, H., Ding, Z., Guo, C., Li, Z., Xia, H. (2008). Research on combination kernel function of support vector machine. *Proceedings - International Conference on Computer Science and Software Engineering*, 1, 838-841.
- Tarihi, G., Selay, D., Yakut, G. (2020). Destek vektör makineleri yardımıyla imalat sanayisinde mali başarısızlık tahminlerinin teknoloji yoğunluğuna göre incelenmesi. *Fen Bilimler Fakültesi Dergisi Journal of Economics and Administrative Sciences*, 4(2), 36-54.
- Toğaçar, M., Ergen, B. (2019). Deep learning approach for classification of breast cancer. **2018 International Conference on Artificial Intelligence and Data Processing**, Elazığ, Türkiye.
- Tong, S., Koller, D. (2001). Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, 10, 45–66.
- Trafalis, T. B., Ince, H. (2000). Support vector machine for regression and applications to financial forecasting. **Proceedings of the International Joint Conference on Neural Networks**, Oklahoma, USA.
- Wang, H., Shi, Y., Zhou, X., Zhou, Q., Shao, S., Bouguettaya, A. (2010). Data Classification Using Support Vector Machine. **Proceedings International Conference on Tools with Artificial Intelligence, ICTAI**, Haryana, India.
- Weston, J., Watkins, C. (1998). Multi-class support vector machine. *Support Vector Machines Applications, Springer International Publishing*, 2, 23–48.
- Widodo, A., Yang, B. S. (2007). Support vector machine in machine condition monitoring. *Mechanical System and Signal Processing*, 21(6), 2560-2574.
- Wold, S., Esbensen, K., Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(3), 37–52.
- Zhang, G., ve Ge, H. (2013). Support vector machine with a Pearson VII function kernel for discriminating halophilic and non-halophilic proteins. *Computational Biology and Chemistry*, 46, 16–22.
- Zhang, L., Zhou, W., Jiao, L., Member, S. (2004). Wavelet support vector machine. *CYBERNETICS*, 34(1), 1083-4419.



ÖZ GEÇMİŞ

Kişisel Bilgiler

Adı Soyadı : Eyyup ADAR

Eğitim Bilgileri

Lisans :
Üniversite : Karadeniz Teknik Üniversitesi
Fakülte : Fen
Bölüm : İstatistik ve Bilgisayar Bilimleri
Mezuniyet Yılı : 2015

Yüksek Lisans :
Üniversite : Van Yüzüncü Yıl Üniversitesi
Enstitü : Fen
Anabilim Dalı : İstatistik
Mezuniyet Yılı : 2023

VAN YÜZÜNCÜ YIL ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
LİSANSÜSTÜ TEZ ORJİNALLİK RAPORU

Tarih 28/02/2023

Tez Başlığı: Destek Vektör Makineler ve Çekirdek Fonksiyonları Üzerine Bir Çalışma

Yukarıda başlığı belirtilen tez çalışmamın, kapak sayfası, giriş, ana bölümler ve sonuç bölümlerinden oluşan toplam 97 (doksanyedi) sayfalık kısmına ilişkin, 28/02/2023 tarihinde şahsım/tez danışmanım tarafından Turnitin adlı intihal tespit programından aşağıda belirtilen filtrelemeler uygulanarak alınmış olan orijinallik raporuna göre tezin benzerlik oranı %7 (yedi) dir.

Uygulanan filtreler aşağıda verilmiştir:

- Kabul ve onay sayfası hariç,
- Teşekkür hariç,
- İçindekiler hariç,
- Simge ve kısaltmalar hariç,
- Gereç ve yöntemler hariç,
- Kaynakça hariç,
- Alıntılar hariç,
- Tezden çıkan yayınlar hariç,
- 7 kelimedenden daha az örtüşme içeren metin kısımları hariç (Limit match size to 7 words)

Van Yüzüncü Yıl Üniversitesi Lisansüstü Tez Orijinallik Raporu Alınması ve Kullanılmasına İlişkin Yönergeyi inceledim ve bu yönergede belirtilen azami benzerlik oranlarına göre tez çalışmamın herhangi bir intihal içermediğini; aksinin tespit edileceği muhtemel durumda doğabilecek her türlü hukuki sorumluluğu kabul ettiğimi ve yukarıda vermiş olduğum bilgilerin doğru olduğunu beyan ederim.

Gereğini bilgilerinize arz ederim.

Tarih ve İmza

Adı Soyadı: Eyyup ADAR

Öğrenci No: 19910001405

Anabilim Dalı: İstatistik

Programı: Tezli Yüksek Lisans

Statüsü: () Yüksek lisans () Doktora

DANIŞMAN

ENSTİTÜ ONAYI

UYGUNDUR

UYGUNDUR

