

T.C.
MUĞLA SITKI KOÇMAN ÜNİVERSİTESİ FEN
BİLİMLERİ ENSTİTÜSÜ

İSTATİSTİK ANABİLİM DALI

SHAPLEY KONSANTRASYON BÖLGESİNDEKİ
GALAKSİLERİN İSTATİSTİKSEL ÖĞRENME
YÖNTEMLERİYLE SINIFLANDIRMASI

YÜKSEK LİSANS TEZİ

NAZLI DENİZ ERGÜÇ

EYLÜL 2019

MUĞLA

MUĞLA SITKI KOÇMAN ÜNİVERSİTESİ
Fen Bilimleri Enstitüsü

TEZ ONAYI

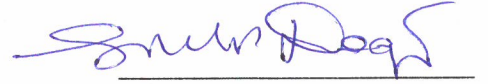
NAZLI DENİZ ERGÜÇ tarafından hazırlanan **Shapley Konsantrasyon Bölgesindeki Galaksilerin İstatistiksel Öğrenme Yöntemleri İle Sınıflandırılması** başlıklı tezinin, 19/09/2019 tarihinde aşağıdaki jüri tarafından İstatistik Anabilim Dalı'nda yüksek lisans derecesi için gerekli şartları sağladığı oybirliği/oyçokluğu ile kabul edilmiştir.

TEZ SINAV JURİSİ

Doç. Dr. Eralp Doğu (**Jüri Başkanı**)

İstatistik Anabilim Dalı,
Muğla Sıtkı Koçman Üniversitesi, Muğla

İmza:



Dr. Öğr. Üyesi Nida GÖKÇE NARİN (**Danışman**)

İstatistik Anabilim Dalı,
Muğla Sıtkı Koçman Üniversitesi, Muğla

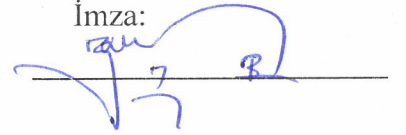
İmza:



Dr. Öğr. Üyesi Volkan OBAN (**Üye**)

Güzel Sanatlar Tasarım ve Mimarlık Anabilim Dalı
İstinye Üniversitesi, İstanbul

İmza:



ANA BİLİM DALI BAŞKANLIĞI ONAYI

Dr. Öğr. Üyesi Aytaç PEKMEZCİ

İstatistik Ana Bilim Dalı Başkan V.,
Muğla Sıtkı Koçman Üniversitesi, Muğla

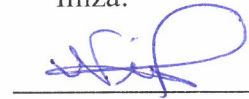
İmza:



Dr. Öğr. Üyesi Nida GÖKÇE NARİN

İstatistik Anabilim Dalı,
Muğla Sıtkı Koçman Üniversitesi, Muğla

İmza:



Savunma Tarihi: 19/09/2019

Tez alıřmalarım sırasında elde ettiėim ve sunduėum tm sonu, dokman, bilgi ve belgelerin tarafımdan bizzat ve bu tez alıřması kapsamında elde edildiėini; akademik ve bilimsel etik kurallarına uygun olduėunu beyan ederim. Ayrıca, akademik ve bilimsel etik kuralları gereėi bu tez alıřması sırasında elde edilmemiř bařkalarına ait tm orijinal bilgi ve sonulara atıf yapıldıėını da beyan ederim.

NAZLI DENİZ ERĐ

19/09/2019

ÖZET

SHAPLEY KONSANSTRASYON BÖLGESİNDEKİ GALAKSİLERİN İSTATİSTİKSEL ÖĞRENME YÖNTEMLERİYLE SINIFLANDIRMASI

Nazlı Deniz ERGÜÇ

Yüksek Lisans Tezi

Fen Bilimleri Enstitüsü

İstatistik Anabilim Dalı

Danışman: Dr. Öğr. Üyesi Nida GÖKÇE NARİN

Eylül 2019, 54 sayfa

Galaksiler, kütle çekim kuvvetiyle bir arada bulunan gaz, toz ve karanlık maddeler ile birlikte farklı boyutlarda, farklı parlaklık derecesinde ve farklı konumlardaki yıldız kümelerinden oluşan sistemlerdir. Evrende milyarlarca galaksi bulunmaktadır ve gelişen teknolojiye paralel olarak her geçen gün yeni galaksi ve gökcisimleri keşfedilmektedir. Her bir galaksinin tek tek incelenmesi oldukça maliyetli olduğundan yeni keşfedilen gök cisimlerinin hâlihazırda bilinen cisimlere benzerliklerini tespit etmek önemli bir problemidir. Galaksileri benzerliklerine göre sınıflandırmak çok sayıda ve farklı özelliklere sahip olan galaksileri daha küçük kompakt gruplar olarak incelemeye imkân sağlar.

Bu amaçla gökyüzünde geniş alanlarda gerçekleştirilen taramalar sonucunda belirlenen galaksileri benzerliklerine göre sınıflandırılan büyük kataloglar üretilmiştir. Galaksi sınıflandırma yöntemleri, veri seti içindeki gizli örüntüyü ortaya çıkararak doğal grupları henüz tespit edilmemiş olan galaksilerin hangi grupta yer alabileceğini tahmin etmek amacıyla kullanılmaktadırlar. Bu sayede gerek araştırmacılara gerekse astronomlara zaman ve maliyet açısından önemli kazançlar sağlanmaktadır.

Bu çalışma da özel olarak Shapley Konsantrasyon bölgesindeki 4215 galaksi, 5 değişken dikkate alınarak sınıflandırılmıştır. Mevcut gökyüzü kataloglarından Abell kataloğu Shapley Konsantrasyon Bölgesindeki bu galaksileri 8 farklı grupta ele almıştır. IDL programlama ile söz konusu bölgedeki galaksilerin hangi Abell grubunda yer aldıkları belirlenmiştir. Doğal grupları tespit edilen galaksiler Weka programı ile istatistiksel ve makine öğrenme algoritmaları kullanılarak sınıflandırılmıştır. Bayes Sınıflandırıcı yöntemlerinden Naive Bayes ve Bayes Ağ, Karar ağaçları yöntemlerinden J48, LMT ve Rasgele Orman algoritmaları, Destek Vektör sınıflandırma yöntemi ve yapay sinir

ağlarından Çok Katmanlı Algılayıcılar kullanılmıştır. Elde edilen sınıflandırma sonuçları doğal gruplarla karşılaştırılarak öğrenme yöntemlerinin tahmin performansları değerlendirilmiştir.

Anahtar Kelimeler: Galaksiler, İstatistiksel Öğrenme, Makine Öğrenmesi, Sınıflandırma, Shapley Konsantrasyon Bölgesi

ABSTRACT

CLASSIFICATION OF GALAXIES IN SHAPLEY CONCENTRATION REGION WITH STATISTICAL LEARNING METHODS

Nazlı Deniz ERGÜÇ

Master of Science (M. Sc.)

Graduate School of Natural and Applied Sciences

Department of Statistic

Supervisor: Assist. Prof. Dr. Nida GÖKÇE NARİN

September 2019, 54 pages

Galaxies are systems consisting of clusters of stars of different sizes, different luminosity, and different positions, together with gas, dust and dark matter combined with gravitational force. There are billions of galaxies in the universe and new galaxies and celestial bodies are being discovered every day in parallel with the developing technology. Since it is very costly to examine each galaxy individually, it is an important problem to identify the similarity of newly discovered celestial bodies to those already known. Classifying galaxies by their similarities allows us to examine galaxies with numerous and different properties into smaller compact groups.

For this purpose, large catalogs, which were classified according to their similarities, were produced as a result of scans carried out in large areas in the sky. Galaxy classification methods reveal the hidden pattern in the data set and use it to estimate which group of galaxies whose natural groups have not yet been identified. In this way, both researchers and astronomers gain significant time and cost savings.

In this study, 4215 galaxies in Shapley Concentration region were classified according to 5 variables. One of the available sky catalogs, the Abell catalog deals with these galaxies in the Shapley Concentration region in 8 different groups. By using IDL programming, it was determined that the galaxies in the region were available in which Abell group was. Galaxies whose natural groups were detected were classified by using Weka program and statistical and machine learning algorithms. Naive Bayes and Bayes Network from Bayes classification methods, J48, LMT and Random Forest algorithms from Decision Trees methods, Support Vector classification method and Multilayer Perceptron from artificial neural networks were used.

The estimation performances of learning methods were evaluated by comparing the results obtained with the natural groups.

Keywords: Galaxies, Statistical Learning, Machine Learning, Classification, Shapley Concentration Region

Çok Değerli Anneme ve Babama

ÖNSÖZ

Tez çalışmalarım boyunca katkılarını ve desteğini esirgemeyen saygıdeğer danışman hocam Dr. Öğr. Üyesi Nida GÖKÇE NARİN'e teşekkürlerimi sunarım.

Tezde kullanılan astronomik bilgileri edinmeme yardımcı olan, desteklerini ve kıymetli zamanlarını esirgemeyen değerli hocalarım Prof. Dr. Aysun AKYÜZ'e ve Doç. Dr. Nazım AKSAKER'e teşekkürü borç bilirim.

Yüksek Lisans eğitimim boyunca her zaman desteğini hissettiğim saygıdeğer hocam Doç. Dr. Nevin GÜLER DİNCER'e teşekkür ederim.

Ayrıca tüm çalışma sürecimde destekleriyle bana güç veren aileme minnettar olduğumu belirtmek isterim.

Bu Tez çalışması Muğla Sıtkı Koçman Üniversitesi Bilimsel Araştırma Projeleri Koordinasyon Birimi tarafından 18/037 kod numaralı “*Shapley Konsanstrasyon Bölgesindeki Galaksilerin İstatistiksel Öğrenme Yöntemleriyle Sınıflandırması*” isimli proje ile desteklenmiştir.

İÇİNDEKİLER

ÖNSÖZ.....	ix
İÇİNDEKİLER	x
ÇİZELGE DİZİNİ	xii
ŞEKİLLER DİZİNİ	xiii
1. GİRİŞ	1
2. LİTERATÜR.....	3
3. GALAKSİLER	10
3.1. Galaksiler.....	10
3.2 Galaksi Sınıflandırması	11
3.2.1. Spiral Galaksiler	12
3.2.2. Eliptik Galaksiler	12
3.2.3. Merceksi Galaksiler	13
3.2.4. Düzensiz Galaksiler	14
4. MATERYAL VE METHOD.....	15
4.1. İstatistiksel Öğrenme ve Makine Öğrenmesi	15
4.2. Öğrenmeye Dayalı Sınıflandırma Yöntemleri.....	16
4.2.1. Bayes Ağ.....	17
4.2.2. Naive Bayes	18
4.2.2. Destek Vektör Makinesi	19
4.2.2.1. Doğrusal Destek Vektör Makinesi.....	19
4.2.2.2. Doğrusal Olmayan Destek Vektör Makinesi.....	21
4.2.3. Çok Katmanlı Algılayıcılar.....	23
4.2.4. Karar Ağaçları.....	25
4.2.4.1. J48 Algoritması.....	27
4.2.4.2. Lojistik Model Ağacı	28
4.2.4.3. Rasgele Orman	29
4.3. Performans Değerlendirme Kriterleri.....	29
4.3.1. Sınıflandırma Başarı Yüzdesi	30
4.3.2. Kappa İstatistiği	30
4.3.3. Karmaşıklık Matrisi	30

4.4. Friedman Testi.....	32
5. UYGULAMA.....	33
5.1. IDL Programlama ile Galaksilerin Doğal Gruplarının Tespiti.....	33
5.2. Weka Sınıflandırma Sonuçları	34
5.2.1. Bayes Ağ.....	36
5.2.2. Naive Bayes	37
5.2.3. Destek Vektör Makinesi	38
5.2.4. ÇKA	39
5.2.5. J48	40
5.2.6. LMT	42
5.2.7. Rasgele Orman.....	42
6. SONUÇLAR ve TARTIŞMA.....	44
KAYNAKLAR	49
ÖZGEÇMİŞ.....	53

ÇİZELGE DİZİNİ

Çizelge 2.1. Morfolojik Galaksi Sınıflandırmasının Literatür Özeti	7
Çizelge 2.1. Morfolojik Galaksi Sınıflandırmasının Literatür Özeti (Devamı)	8
Çizelge 2.1. Morfolojik Galaksi Sınıflandırmasının Literatür Özeti (Devamı)	9
Çizelge 4. 1. Çekirdek Fonksiyonları.....	23
Çizelge 4. 2. Karmaşıklık Matrisi	31
Çizelge 5.1. Gruplar Arasındaki Öklid Uzaklıkları	34
Çizelge 5.2. Weka formatına Dönüştürülmüş Veri Seti.....	35
Çizelge 5. 3. Bayes Ağ Yöntemi için Karmaşıklık Matrisi	37
Çizelge 5. 4. Naive Bayes Yöntemi için Karmaşıklık Matrisi	37
Çizelge 5. 5. Çekirdek Fonksiyonlarının Uygulama Sonuçları	38
Çizelge 5. 6. Destek Vektör Sınıflandırma Yöntemi için Karmaşıklık Matrisi.....	38
Çizelge 5. 7. Çok Katmanlı Ağ Sınıflandırma Yöntemi için Karmaşıklık Matrisi.....	40
Çizelge 5. 8. J48 Sınıflandırma Yöntemi için Karmaşıklık Matrisi	41
Çizelge 5. 9. Sınıflandırma Yöntemi için Karmaşıklık Matrisi.....	42
Çizelge 5.10. Rasgele Orman Sınıflandırma Yöntemi için Karmaşıklık Matrisi	43
Çizelge 6. 1. Sınıflandırma Yöntemlerinin Performans Karşılaştırması.....	44
Çizelge 6. 2. Friedman Test Sonuçları.....	45

ŞEKİLLER DİZİNİ

Şekil 3. 1. Hubble Çatal Diyağramı.....	11
Şekil 3. 2. Spiral Galaksi Örneği.....	12
Şekil 3. 3. Eliptik Galaksi Örneği.....	13
Şekil 3. 4. Düzensiz Galaksi Örneği.....	14
Şekil 4. 1. Bayes Ağ Modeli	18
Şekil 4. 2. Doğrusal DVM Örneği.....	20
Şekil 4. 3. Doğrusal Olmayan DVM Örneği.....	21
Şekil 4. 4. Çok Katmanlı Algılayıcı Modeli	24
Şekil 4. 5. Örnek Karar Ağacı Modeli	26
Şekil 5. 1. Abell gruplarında bulunan galaksi gözlem sayılar.....	35
Şekil 5. 2 . Weka platformunda sınıflandırma aracı	36
Şekil 5. 3. Shapley konsanstrasyon bölgesine ait çok katmanlı sinir ağı yapısı	39
Şekil 5. 4 . J48 yönteminden elde edilen modelin bir parçası.....	41
Şekil 6. 1. Yöntemler arasındaki anlamlı farklılığın grafiksel gösterimi	46

SEMBOLLER ve KISALTMALAR

P(A)	A Olayının Gerçekleşme Olasılığı
P(B)	B Olayının Gerçekleşme Olasılığı
P(A/B)	B Olayı Biliniyorken A Olayının Gerçekleşme Olasılığı
P(B/A)	A Olayı Biliniyorken B Olayının Gerçekleşme Olasılığı
w	Ağırlık Vektörü
b	Yanlılık Katsayısı
X_{net}	Yapay Sinir Ağı Modelindeki Girdi Değerleri
$f(x_i)$	Eşik fonksiyon
ε_t	Modelin hatası
S	Örnek uzayı
P_T	Gözlemlenen Uyumların Toplam Olasılığı
P_e	Gözlemlerin Tesadüfi Uyumunun Olasılığı
a	Wilcard değeri
SS	Kareler Toplamı
df	Serbestlik Derecesi
MS	Ortalama Kare Değeri
Chi-sq	Test İstatistiği
H₀	Sıfır Hipotez
H_a	Alternatif Hipotez
$\phi(x_j)$	Çekirdek Fonksiyon
$\vec{x}_i, \vec{x}_j$	Girdi Vektörleri
YSA	Yapay sinir ağıları
DVM	Destek Vektör Makinesi

LMT	Lojistik Model Ağacı
J48	Karar Ağacı Sınıflandırıcısı
RO	Rasgele Orman
KA	Karar Ağacı
ÇKA	Çok Katmanlı Algılayıcı
Kİ	Kappa İstatistiği
KM	Karmaşıklık Matrisi
TP	True Positive
FN	False Negative
FP	False Positive
TN	True Negative
BY	Başarı Yüzdesi
R.A	Boylam
Dec	Enlem
Mag	Parlaklık
V	Hız
SİGV	Hızdaki Sapma Değeri

1. GİRİŞ

Galaksiler, kütle çekim kuvvetiyle bir arada bulunan yıldızlardan, gaz, toz ve karanlık maddelerden oluşan devasa sistemlerdir. Farklı boyutlarda, farklı parlaklık derecesinde ve farklı konumlardaki yıldız kümelerinden oluşan galaksiler aynı zamanda evrenin geçmişi hakkında da araştırmacılara bilgiler sunmaktadır. Son yıllarda teknolojiye gelişmelere paralel olarak gök cisimlerine, yıldızlara ve galaksilere ait çok sayıda gözlemler gerçekleştirilmekte ve yeni gökcisimleri tespit edilmektedir. Gözlemler sonucunda elde edilen gökyüzü cisimlerinin özellikleri benzerliklerine göre sınıflandırılmaktadır.

Galaksi sınıflandırması astronomik veri analizinde oldukça önemli bir yer tutmaktadır. Galaksiler morfolojik ve spektral sınıflandırma olarak iki farklı şekilde sınıflandırılmaktadır. Geniş alanlarda gerçekleştirilen taramalar sonucunda belirlenen galaksilerin benzerliklerine göre sınıflandırılmasıyla araştırmacılara daha kompakt bilgiler sunan büyük kataloglar üretilmiştir.

Galaksi sınıflandırması ve kataloglama süreci oldukça zaman alan ve maliyeti yüksek çalışmalardır. Galaksi sınıflandırması için genelleme başarısı yüksek olan makine öğrenmesi ve istatistiksel öğrenme algoritmaları kullanılabilir. Bu öğrenme algoritmalarının genel mantığı belirli bir veri kümesi içinde yer alan gizli bilgiyi ortaya çıkarmaya dayanmaktadır. Algoritmalar belli bir eğitim setini kullanarak bir problemin çözümü için eğitilir ve ardından daha önce karşılaşmadığı test verileri için tahminde bulunurlar. İstatistiksel öğrenme ve makine öğrenmesi yaklaşımları ile oluşturulan galaksi kümelerinin mevcut kataloglardaki kümelere uyumlu olduğunun gösterilmesi zaman ve maliyet açısından astronomlara ve araştırmacılara kazanç sağlayabilir.

Bu amaçla, bu çalışmada danışmanlı öğrenmeye dayalı sınıflandırma yöntemleriyle galaksi sınıflandırması yapılmış ve yöntemlerin sınıflandırma performansları karşılaştırılmıştır. Hedeflenen amaç doğrultusunda tez çalışmasında Shapley Konsantrasyon bölgesinde yer alan 4215 galaksi, boylam (R.A), enlem (Dec.), parlaklık

(Mag.), hız (V), ve hızdaki sapma (SIGV) deęişkenleri dikkate alınarak incelenmiştir. Shapley Konsantrasyon bölgesindeki galaksiler George O. Abell tarafından oluşturulan Abell kataloęundaki 8 farklı grupta yer almaktadır (Abell, 1957, Abell vd., 1989). IDL programı kullanılarak hangi galaksinin hangi Abell grubunda yer aldığı tespit edilmiş ve doęal grupları tespit edilen galaksilere Bayes Ağları, Destek Vektör Makinesi (DVM), Karar Ağaçları ve Yapay Sinir Ağlarına dayalı sınıflandırma yöntemleri uygulanmıştır. Kullanılan sınıflandırma yöntemlerinin performansları Abell kataloęuna uyumlarına göre deęerlendirilmiştir.

2. LİTERATÜR

Literatürde galaksi sınıflandırmasına ilişkin pek çok çalışma bulunmaktadır. Çalışmaların büyük bir kısmında galaksiler ya morfolojilerine göre ya da yıldız spektrallerine göre sınıflandırılmıştır. Tez çalışmasında yalnızca morfolojik sınıflandırma ele alındığı için bu bölümde morfolojik sınıflandırmaya ilişkin öne çıkan literatür çalışmalarına yer verilmiştir.

Galaksi sınıflandırması ilk kez Edwin Hubble tarafından 1926'da yapılmıştır. Hubble (1926) çalışmasında galaksileri Eliptik, Sarmal (spiral) ve merceksi olmak üzere üç ana grupta sınıflandırmıştır. Hubble düzeni olarak adlandırılan bu sınıflandırma yaygın bir morfolojik sınıflandırma biçimi olarak kabul edilmiştir. Bu sınıflandırma biçimine ek olarak Sparke ve Gallagher tarafından hiçbir gruba dâhil edilemeyen düzensiz galaksiler tanımlanmıştır (Sparke Gallagher, 2007).

Dressler (1980) çalışmasında galaksilerin morfoloji-yoğunluk ilişkisini incelemiş ve galaksilerin küme merkezine uzaklıklarına göre değil de galaksi yoğunluklarına göre kümелendiğini göstermiştir. Spiral ve düzensiz galaksilerin yoğunluklarına göre dağınık bir şekilde kümелendiğini ve bu durumun morfoloji ile yerel galaksi yoğunluğu arasındaki ilişkiden kaynaklandığını ifade etmiştir.

Miller ve Coe (1996) çalışmasında, Kohonen kendi kendini düzenleme haritasına (SOM) dayanan bir sınıflandırıcı sunulmuştur. SOM'un eğitiminde kullanılan eğitim seti, bir grup görüntü üzerinden elde edilen özelliklerin elle sınıflandırılması ile oluşturulmuştur. Çalışma sonunda %98 doğruluk oranıyla galaksi sınıflandırılması gerçekleştirilmiştir.

Goderya ve Lolling (2001), galaksilerin biçimsel özelliklerini kullanarak geometrik şekil sınıflandırması için yapay sinir ağına dayanan otomatik bir galaksi sınıflandırması yöntemi önermişlerdir.

Lotz vd. (2004) yılında yaptıkları çalışmada galaksi morfolojisini ölçmek için Gini Katsayısı ve M(20) olarak adlandırılan parametrik olmayan iki yöntem önermişlerdir. Bu çalışmada, 148 parlak Hubble tipi galaksinin kümelenme ölçümleri incelenmiştir. Parlak çekirdekli galaksilerin Gini katsayısının yüksek olduğu, çoklu çekirdekleri olan

galaksilerin ve parlak gel-git kuyrukları olan galaksilerin ise ikinci dereceden nispi dağılıma ($M(20)$) sahip olduğu gösterilmiştir.

Schneider (2006) çalışmasında galaksilerin yüzey parlaklıklarını modelleyerek morfolojik galaksi sınıflandırması için yeni bir yöntem önermiştir. Bu yöntemle göre Vaucouleurs profili olarak tanımlanan galaksilerin eliptik bir formda olduğu, üstel yüzey parlaklık profiline sahip bir galaksinin ise spiral galaksi olduğu ifade edilmiştir.

Siddhartha vd. (2007) çalışmasında makine öğrenme algoritmaları kullanılarak morfolojik galaksi sınıflaması yapılmıştır. Bu çalışmada Zolt Fre galaksi kataloğundan alınan 119 görüntü işlenmiş ve yine aynı katalogdan elde edilen sınıflandırma etiketleri baz alınarak Destek Vektör Makineleri, Rasgele Orman ve Naive Bayes algoritmaları kullanılarak galaksiler sınıflandırılmıştır. Rasgele Orman algoritmasının performansının daha yüksek olduğu görülmüştür.

Bailin ve Harris (2008) çalışmasında geniş kapsamlı gökyüzü araştırmalarını içeren büyük dijital gökyüzü anketi (Sloan Digital Sky Survey) verilerini inceleyerek erken tip, geç tip ve orta tip olmak üzere üç tip galaksi tanımlanmıştır. Erken tip galaksilerin kırmızı, yoğunlaşmış ve yuvarlak özelliklerde olduğu, geç tip galaksilerin düşük yoğunluklu disk şeklinde ve mavi olduğu, orta tip galaksilerin ise orta büyüklükte yoğunluğa sahip ve disk şeklinde olduğu tespit edilmiştir. Elde edilen sonuçların doğruluğu Millennium Galaksi Kataloğundaki galaksi verilerin sınıflandırma sonuçlarıyla karşılaştırılarak ölçülmüştür.

Gauci vd. (2010) galaksi türlerini ve galaktik cisimleri ayırt etmek için karar ağacı öğrenme algoritmalarını ve bulanık çıkarım sistemlerini kullanmışlardır. Sonuçlar Galaxy Zoo kataloğu ile karşılaştırılmış ve en iyi uyuma sahip yöntem Rasgele Orman olarak tespit edilmiştir.

Marin vd. (2013) çalışmasında Schmidt kamerasından 50 yıllık bir periyodu kapsayan astronomik tabakaları içeren, INAOE veri setini kullanmıştır. Bu veri seti 24 astronomik katmanda yer alan 152 galaksiyi içermektedir. Ancak bu sayı yetersiz olduğu için orijinal gözlemler üzerinden geometrik dönüşümlerle yapay örnekler üretilmiştir. Bu örneklerin Naive Bayes ve Rasgele Orman algoritmaları ile morfolojik olarak sınıflandırılması sonucunda Rasgele Orman algoritmasının performansının daha yüksek olduğu görülmüştür.

Gauthier vd.(2016) galaksileri spiral, eliptik, yuvarlak, disk veya diğer olarak sınıflandırmıştır. Bu sınıflandırma işlemini Karar ağaçları, Adaboost sınıflandırıcıları, RBF (Radyal Temel Çekirdek) çekirdek, K-En Yakın Komşu Algoritması, Destek Vektör Makinesi ve Rasgele Orman yöntemleriyle yapmış ve en iyi performansı % 94 doğruluk ile Rasgele Orman yöntemi sağlamıştır. Ayrıca, galaksilerin her bir sınıfla ilişkilendirilme ihtimalini tahmin etmek için de regresyon kullanılmıştır

Remya ve Mohan (2016), galaksi sınıflandırması için yapay sinir ağı yöntemini kullanmışlardır. Sinir ağı Galaxy Zoo projesi'nden elde edilen verilerle eğitilmiş ve galaksilerin hangi morfolojik yapıya sahip olduğu, doğrudan ham piksel verilerinden tahmin edilmiştir.

Selim vd. (2016) Zsolt Frei kataloğundaki 113 görüntüden oluşan bir eğitim seti ve 20 görüntüden oluşan test setini negatif olmayan matris faktörizasyonu ile % 93 doğrulukla sınıflandırmışlardır.

Dobrycheva vd. (2017), renk indekslerine göre makine öğrenmesi kullanarak galaksileri eliptik, spiral ve düzensiz olarak üç grupta sınıflandırmıştır. Naive Bayes, Rasgele Orman ve Destek Vektör sınıflandırma yöntemleri arasından Rasgele Orman en yüksek doğruluk oranına sahip olmuştur.

Paul vd. (2018) çalışmasında yaklaşık 3.0000.000 SDSS (tüm gökyüzü kataloğu) galaksinin kırmızıya kayma değerlerine göre bir katalog oluşturmuşlardır. Bu çalışmanın temel amacı, galaksi morfolojisi ve Fotometrik kırmızıya kaymayı hesaplamak için makine öğrenmesini ve değişken seçim algoritmalarını test etmek, belirli bir galaksi popülasyonu için temel olarak uygulamaktır. Basit Doğrusal Regresyon, Çok katmanlı Algılayıcı, MP5, Zror, Karar Tablosu ve Rasgele Orman yöntemleri ile görüntü işleme yapılmıştır. ANOVA ile değişken seçimi yapılmıştır. RO yöntemi diğer yöntemlere göre en başarılı yöntem olarak bulunmuştur.

Cheng vd. (2019) çalışmasında galaksi sınıflandırması için görüntüleme verilerini kullanarak en uygun makine öğrenme yöntemlerini belirlemeyi amaçlamışlardır. Galaxy Zoo 1 projesinden (GZ1) görsel sınıflandırmalarla birleştirilen Dark Energy Survey (DES) verilerini kullanarak Konvolüsyonel Sinir Ağı (CNN), K-en yakın komşu, Lojistik Regresyon, Destek Vektör Makinesi ve Sinir Ağları yöntemlerini kullanmışlardır. En iyi

performansı CNN göstermiştir. Bu yöntem spiral ve eliptik galaksileri 0.94 doğrulukla ayırt etmiştir.

Galaksi kataloglama tüm gökyüzündeki galaksileri türlerine, benzer özelliklerine ve konumlarına göre gruplandırma işlemidir. Kataloglar tüm gökyüzündeki galaksilerin bilinen tüm özelliklerini içeren temel kaynaklardır. Galaksi sınıflandırması ise daha küçük bölgedeki galaksilerin belirli özellikleri dikkate alınarak benzer grupların oluşturulması işlemidir.

Abell (1957) çalışmasında Kuzey yarımküredeki galaksileri zenginlik, yoğunluk, mesafe, galaktik enlem kriterlerine göre kataloglamıştır. Abell (1988) kataloğu ise Güney yarımküre 'deki 1.361 kümeyi de içermektedir.

Driver vd. (2005) çalışmasında galaksileri yoğunluğuna ve parlaklığına göre kataloglamıştır

Bu çalışma da istatistiksel ve makine öğrenmesi yöntemleri kullanılarak galaksi sınıflandırması yapılmıştır. Literatürde yer alan morfolojik sınıflandırma yaklaşımları ve galaksi kataloglarının bazıları Çizelge 2.1' de özetlenmiştir. Bu çalışmada özel olarak ele alınan Shapley Konsantrasyon bölgesi, Harlow Shapley ve arkadaşları tarafından 1930'da keşfedilmiştir. Shapley vd., Güney Afrika'nın Bloemfontein bölgesinde 24 inçlik Bruce teleskobundan elde ettikleri fotoğraf plakalarını kullanarak kuzey gökyüzündeki galaksiler üzerine bir çalışma yapmışlardır (Shapley 1930). Ayrıca 1932'de çalışmasında güney gökyüzünün üçte birinde yer alan ve parlaklığı 18. görünür parlaklığından daha parlak olan 76.000 galaksiyi keşfedilmiştir (Shapley 1932). Bu 76.000 galaksinin kırmızıya kayma özelliğine göre Drinkwater tarafından belirlenen 4215 galaksi gözlemi bu tez çalışmasında ele alınmıştır (Drinkwater, 2004).

Bu çalışma için özel olarak Shapley Konsantrasyon bölgesi kullanılmıştır. Bu bölgenin tercih edilmesinin nedeni iç içe girmiş çok fazla galaksi yapıları mevcuttur. Bu çalışmada incelenecek olan Shapley Konsantrasyon bölgesindeki galaksilere ilişkin veri seti (<http> 1) ve Abell Kataloğu ilişkin veriler Vizier veri tabanından (<http> 2) alınmıştır. İstatistiksel öğrenme ve makine öğrenme yaklaşımlarıyla elde edilen grupların Abell ile uyumlu olduğu gösterilmesi araştırmacılara avantaj sağlayacaktır.

Çizelge 2. 1 Morfolojik Galaksi Sınıflandırmasının Literatür Özeti

Yazar	Yöntem	Değişkenler	Sonuç
Hubble 1926	Morfolojik Sınıflandırma	Fotografik görüntülerdeki formlar	Galaksiler eliptik, sarmal (spiral) ve merceksi olmak üzere ilk kez üç ana grupta sınıflandırılmıştır.
Vaucouleurs (1959)	Morfolojik Sınıflandırma	Galaksi Morfolojileri	Hubble düzeninde tanımlanmış eliptik, sarmal ve düzensiz sınıfları kendi içerisinde de sınıflara ayırmıştır.
Sandage (1961)	Morfolojik Sınıflandırma	Galaksi Morfolojileri Alt sınıflar	Önceden bilinen Hubble sınıflarının alt sınıflara ayrılmasına yönelik çalışmalar yapmıştır.
Lotz vd. (2004)	Gini Katsayısı ve M(20)dağılımı	Parlak Hubble tipi galaksiler	Parlak çekirdekli galaksilerin Gini katsayısının yüksek olduğu, çoklu çekirdekleri olan galaksilerin ve parlak gel-git kuyukları olan galaksilerin ise M(20) dağılımına sahip olduğunu göstermiştir.
Dressler (1979)	Görüntü İşleme, Yoğunluğa dayalı sınıflandırma	Galaksi yoğunlukları, Küme merkezleri	Galaksi morfolojisi ile galaksi yoğunluğu arasındaki ilişkiyi incelemiştir.
Siddhartha vd. (2007)	Naive Bayes Rasgele Orman Destek Vektör Temel Bileşen Analizi	Görüntü Pikselleri	119 görüntü işlenmiş ve sınıflandırma sonuçlarından Rasgele Orman algoritmasının performansının daha iyi olduğu görülmüştür.
Marin vd.(2013)	Naive Bayes Rasgele Orman	24 Astronomik katmandaki galaksiler	152 galaksi ele alınmıştır. Yapay örnekler üretilmiş ve Rasgele Orman algoritmasının Naive Bayes algoritmasından daha yüksek bir performans sergilediği ve yapay örneklerin sınıflandırma performansında katkı sağladığını göstermiştir.
Miller ve Coe (1996)	SOM	Görüntü Pikselleri	Yıldızları ve galaksileri ayırt edebilmek için morfolojik sınıflandırma yapılmıştır. Çalışma sonunda %98 doğruluk sağlanmıştır.

Çizelge 2. 2 Morfolojik Galaksi Sınıflandırmasının Literatür Özeti (Devamı)

Yazar	Yöntem	Değişkenler	Sonuç
Bailin ve Harris (2008)	Görüntü İşleme	Sloan Digital Sky Survey) galaksi verileri	Erken tip, geç tip ve orta tip olarak 3 tip galaksi tanımlanmıştır. Sonuçlar Millennium Galaksi Kataloğundaki verilerin sınıflandırma sonuçlarıyla karşılaştırılarak ölçülmüştür
Gauci vd.(2010)	Karar Ağaçları ve Bulanık Mantık	Fotometrik Özellikleri	SDSS verileri ile sınıflandırma yapılmış, sonuçlar Galaxy Zoo kataloğu ile karşılaştırılmıştır. Yöntemlerden Rasgele Orman en iyi sonucu vermiştir.
Gauthier vd.2016	RBF çekirdek Rasgele Orman AdaBoost KNN	Galaksi görüntü formları	Galaksileri spiral, eliptik, yuvarlak, disk veya diğer olarak sınıflandırılmış ve random forest yöntemi en iyi performansı sağlamıştır.
Dobrycheva vd 2017	Renk endeksi Naive Bayes, Rasgele Orman Destek Vektör	Mutlak Büyüklükler, Renk endexleri	Renk endekslerine göre galaksilerin biçimleri tanımlanmış ve Naive Bayes, Rasgele Orman ve Destek Vektör yöntemleri uygulanmıştır. Rasgele Orman yöntemi en yüksek doğruluğu sağlamıştır.
Remya G, Anuraj Mohan 2016	Derin öğrenme Evrişimli sınır ağı	Görüntü pikselleri	Galaxy Zoo projesinden elde edilen veriler eğitilmiş ve galaksi morfolojisinin çeşitli yönlerini doğrudan ham piksel verilerinden tahmin edebilmiştir.
I. M.Selim vd. 2016	Negatif Olmayan Matris Faktörizasyonu	Görüntü Pikselleri	Zsolt Frei kataloğundaki 113 görüntüden oluşan bir eğitim seti ve 20 görüntüden oluşan test seti sınıflandırılmış 93% doğruluk sağlamıştır.
Goderya ve Lolling 2001	Geometrik şekil özellikleri ve piksel görüntü sınıflandırma	Galaksilerin geometrik şekilleri ve görüntü pikselleri	Geometrik şekil sınıflandırmasında biçimsel özelliklerine ve yapay sınır ağına dayanan bir otomatik galaksi sınıflandırmasının geliştirilebileceğini göstermektedir.

Çizelge 2. 3 Morfolojik Galaksi Sınıflandırmasının Literatür Özeti (Devamı)

Yazar	Yöntem	Değişkenler	Sonuç
Schneider (2006)	Yüzey parlaklıklarıyla morfolojik sınıflandırma	Yüzey parlaklıkları	Vaucouleurs profili olarak tanımlanan modelde galaksinin eliptik bir formda olduğu, üstel yüzey parlaklık profiline sahip bir galaksinin ise spiral galaksi olduğu ifade edilmiştir.
Abell (1957), Abell (1989)	Galaksi Kataloglama	Zenginlik Yoğunluk Uzaklık Galaktik Enlem	Abell 1957 çalışmasında Kuzey yarımkürede galaksileri Zenginlik, Yoğunluk, Uzaklık Galaktik Enlem kriterlerine göre kataloglamıştır. Abell 1989 çalışmasında da Güney Yarımküredeki 1.361 kümeyi de dâhil etmiştir.
Driver S.P vd. (2005)	Galaksi Kataloglama	Yoğunluk ve Parlaklığı	Driver S.P et all çalışmasında galaksileri yoğunluğuna ve parlaklığına göre kataloglamıştır.
Zsolt Frei vd. (1996)	Galaksi Kataloglama	Dijital Görüntüler	Zsolt Frei et all.113 galaksinin görüntülerinden oluşan dijital bir katalog oluşturmuşlardır.

3. GALAKSİLER

3.1 Galaksiler

Gökyüzü araştırmaları geçmişten günümüze yoğun ilgi ve merak uyandıran bir alandır. Geçmişten beri insanlık, evrenin oluşumunu anlamak için evreni oluşturan cisimleri, olayları anlamaya çalışmaktadır. Bu amaçla geliştirilen teleskoplar ve uzay araçları yardımıyla evren hakkında bilgi toplanmaktadır. Evrenin temel taşı olan galaksiler kütle çekim kuvvetiyle bir arada bulunan yıldızlar, karanlık maddeler ile yıldızlar arası gaz ve toz bulutlarından oluşmaktadırlar. 20. yy başlarında insanlık tarihi evren hakkında sadece Samanyolu galaksisi ile ilgili bilgilere sahip iken 1920'li yıllarda Edwin Hubble çalışmasında samanyolu galaksisi dışında da galaksilerin olduğunu duyurmuştur (Hubble 1926).

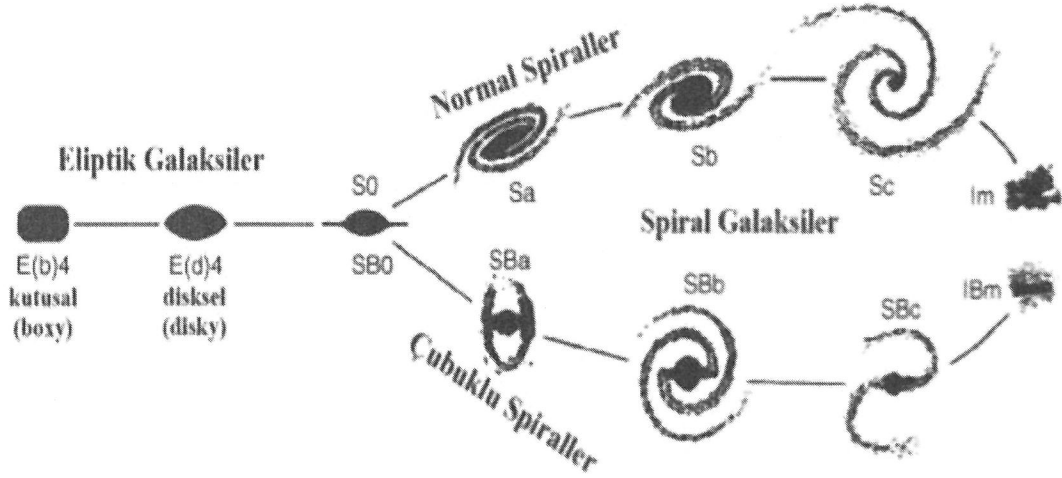
Galaksilerin içerisinde bulunan milyarlarca yıldızların her biri tıpkı insanlar gibi doğar ve ölür. Bu doğum ve ölüm süreleri on binlerce yılı aşabilir. Galaksiler evrende tek başına olmaktan ziyade genellikle kümeler halindedir. Bu kümeler içerisinde galaksiler zaman zaman çarpışabilir. Çarpışmalar yeni yıldız oluşumlarına yol açabilir. Hemen hemen tüm galaksilerin merkezinde büyük kütleli karar delikler mevcuttur (Schilling, 2014).

Günümüze kadar geniş gökyüzü tarama projeleriyle çok sayıda yeni tip galaksiler, yıldızlar ve gök cisimleri tespit edilebilmiştir ve çalışmalar halen devam etmektedir. Galaksiler farklı biçimsel özelliklere sahip olabilirler. Farklı biçimlere ve parlaklıklara sahip galaksilerin evrende dağılımları ve yaydıkları ışınlamalar araştırmacılara evren hakkında bilgi vermektedir. Bu nedenle benzer özelliklere sahip galaksilerin tespit edilmesi açısından galaksi sınıflandırması önemli bir problem olarak incelenmektedir.

3.2. Galaksi Sınıflandırması

Galaksiler evrende homojen bir şekilde dağılmamışlardır. Galaksi grupları veya kümeleri evreni anlamak adına bilim insanları için bir laboratuvar konumundadır. Çeşitli morfolojik yapılar galaksilerin evrimsel süreçleri hakkında bilgiler vermektedir.

Her bir galaksinin tek tek incelenmesi zaman ve maliyet açısından zorlu süreçlerdir. Galaksi sınıflandırması bu açıdan astronomik veri analizinde önemli bir yer tutmaktadır. Galaksi sınıflandırması morfolojik ve spektral sınıflandırma olmak üzere ikiye ayrılır. Morfolojik sınıflandırma galaksilerin biçimsel özelliklerini dikkate alırken, spektral sınıflandırma ise yıldız tayflarını dikkate alarak sınıflandırma yapmaktadır.



Şekil 3.1. Hubble Çatal Diyagramı (Kormendy ve Bender, 1996)

Şekil 3.1’de Hubble Çatal diyagramı verilmiştir. İlk kez Hubble tarafından gerçekleştirilen morfolojik sınıflandırmada galaksilerin biçimsel özelliklerine göre farklılıklar gösterildiği ortaya konulmuştur. Günümüzde değişiklikler gösterse de halen kullanılan Hubble Çatal diyagramında galaksiler spiral, merceksi ve eliptik olmak üzere 3 farklı şekilde sınıflandırılmıştır. Biçimsel olarak herhangi bir gruba dâhil edilemeyen düzensiz galaksiler ise daha sonradan bu sınıflandırma sistemine dâhil edilmiştir (Sparke ve Gallagher, 2007).

3.2.1 Spiral Galaksiler

Spiral galaksiler, kolları ve merkezi şişkin bir yapıdan oluşan disk şeklindeki galaksi biçimidir. Spiral galaksiye ilişkin bir örnek Şekil 3.2’de verilmiştir.



Şekil 3. 2. Spiral galaksi örneği ([http 3](http://3))

Kendi içerisinde normal (S) ve çubuklu (SB) olarak iki alt sınıfa ayrılır. Bu sınıflar da spiral galaksilerin merkezi şişkin yapısına ve parlaklık yapılarına göre a, b, c harfleri ile gösterilen alt sınıflara ayrılmaktadır. Kollarında çoğunlukla sıcak ve parlak tip yıldızlar bulunmaktadır (Schneider, 2006).

3.2.2 Eliptik Galaksiler

Evrende en çok görülen galaksi türlerinden olan eliptik galaksiler küre şeklinde olup spiraller gibi kolları yoktur. İçerisinde yaşlı yıldız sayısı oldukça fazladır. Evrende ki en büyük galaksilerin eliptik galaksi türleri olduğu düşünülmektedir. Bu galaksiler eliptik formlarına göre alt sınıflara ayrılmaktadır. Bu alt sınıflar yuvarlaklık ve basıklık ölçülerine göre belirlenmektedir.

Yuvarlaklıktan basıklığa doğru E0’dan E7’ye kadar 8 farklı alt gruba ayrılmaktadır. E0 olarak adlandırılan galaksiler yuvarlak bir forma sahip iken E7 olarak adlandırılan

galaksiler en basık olan eliptik galaksi türüdür. (Schneider, 2006). Şekil 3.3'de eliptik galaksi örneği verilmiştir.



Şekil 3. 3. Eliptik galaksi örneği (http 4)

3.2.3 Merceksi Galaksiler

Merceksi galaksiler disk şekli ile spiral galaksilere benzer olup S0 olarak adlandırılmıştır. Bu galaksilerin spiral galaksilerden farkı spiral kollarının olmamasıdır. Çok fazla yaşlı yıldız barındıran merceksi galaksilerde gaz ve toz bulutları daha az görülmektedir. Bazı merceksi galaksiler çubuk yapısının olup olmamasına göre alt sınıflara ayrılırlar. Çubuklu merceksi galaksiler S0, çubuklu yapı görülmeyen merceksi galaksiler SB0 olarak adlandırılır. (Schneider 2006).

3.2.4 Düzensiz Galaksiler

Belirli bir morfolojik şekle sahip olmayan farklı şekillerde görülebilen düzensiz galaksiler Irr sembolü ile gösterilmektedir. Bu galaksiler karmaşık yapılara sahip olan galaksilerin çarpışması veya birleşmesi ile oluşmaktadırlar. Düzensiz galaksilerde yeni oluşan genç yıldızlar, gaz ve toz bulutları yoğun olarak görülmektedir. (Hubble,1926; Buta, 2011) Düzensiz galaksi örneği Şekil 3.4'te görülmektedir.



Şekil 3. 4. Düzensiz galaksi örneği ([http 5](http://5))

4. MATERYAL VE METHOD

4.1 İstatistiksel Öğrenme ve Makine Öğrenmesi

Gelişen teknolojiden, üretilen yüksek teknolojiyle uyumlu teleskoplardan, gökyüzü araştırmaları için gönderilen uzay araçlarından elde edilen veri miktarı her geçen gün artmaktadır. Bu da gökyüzü araştırmacıları için geniş veri kaynakları sağlamaktadır. Bu verilerden anlamlı bilgiler çıkarılabilmesi evrenin anlaşılabilmesi açısından araştırmacılar için önemlidir.

Öğrenme yöntemleri büyük veri kümeleri içerisinde yer alan gizli bilgiyi / örüntüyü ortaya çıkarmayı amaçlayan genel algoritmalar bütünüdür. Makine öğrenmesi ve istatistiksel öğrenme temelde aynı problemlerin çözümü için kullanılsa da farklılaştığı noktalar vardır. İstatistiksel öğrenme, kural tabanlı çıkarımlara dayanır ve varsayımlar altında çalışır. Makine öğrenmesinde varsayımlar ve varsayım kontrolleri daha az olduğundan daha esnek bir yapıya sahiptir. Aykırı gözlem ve eksik gözlem gibi durumlardan daha az etkilenmektedir. Bu iki yöntem arasındaki en temel fark istatistiksel öğrenme de açıklanabilir modeller elde edilirken makine öğrenmesinde açıklanabilirliği daha düşük modeller oluşturulur. İstatistiksel öğrenmede model yorumlaması ile sonuca ulaşırken makine öğrenmesinde ise tahmin yakınlığı ile sonuç elde edilir (Aydın, 2011).

Öğrenme algoritmaları veri ile beslenerek kendi mantığını oluşturmaktadır. Veri seti genellikle hold-out metoduna göre %70 eğitim ve %30 test olmak üzere ikiye ayrılmaktadır. Eğitim seti üzerinden eğitim gerçekleştirilir. Veri setindeki örüntünün öğrenilerek model oluşturulması hedeflenmektedir. Test aşamasında ise oluşturulan bu model test edilerek öğrenme yönteminin başarısı belirlenmektedir.

Öğrenme algoritmaları veri setinden bilgiyi elde etme yaklaşımlarına göre temelde iki gruba ayrılmaktadır: danışmanlı (supervised) ve danışmansız (unsupervised) öğrenme. Ele alınan probleme ilişkin sadece girdi verileri biliniyor, çıktı verileri bilinmiyorsa danışmansız öğrenme, girdi verilerine karşılık hedef değerler biliniyor ise danışmanlı öğrenme yöntemleri kullanılmaktadır. Sınıflandırma ve regresyon problemleri doğası

gereği danışmanlı öğrenme yaklaşımları kullanılarak çözülür. Kümeleme problemlerinin çözümü içinde danışmansız öğrenme algoritmaları kullanılır (Han vd, 2006). Öğrenme algoritmaları belli bir eğitim setini kullanarak bir problemin çözümü için eğitildiğinde aynı problem için daha önce karşılaşmadığı test verileri için de tahmin yapabilme gücüne sahiptir (James vd. 2013).

4.2 Öğrenmeye Dayalı Sınıflandırma Yöntemleri

Öğrenmeye dayalı sınıflandırma yöntemleri, eğitim setindeki sınıflara ilişkin özellikleri öğrenerek test setindeki verilerin doğal sınıflarını tahmin etmek için kullanılmaktadır. İstatistiksel öğrenmeye dayalı sınıflandırma yöntemlerinden Bayes sınıflandırıcılar, Destek Vektör Makineleri ve Karar Ağaçları kullanılmıştır. Makine öğrenmesi sınıflandırma yöntemlerinden olan Yapay Sinir Ağlarına dayalı Çok Katmanlı Algılayıcılar kullanılmıştır. Özel olarak bu çalışmada, galaksileri sınıflandırması amacıyla Bayes sınıflandırıcılardan Bayes Ağları ve Naive Bayes, Destek Vektör Makinesi (DVM) ve Karar Ağaçlarından J48, LMT ve Rasgele Orman sınıflandırma yöntemi, Yapay Sinir Ağlarından ise Çok Katmanlı Ağ sınıflandırma (ÇKA) yöntemi kullanılmıştır.

4.2.1 Bayes Sınıflandırıcılar

Bayes sınıflandırıcılar Bayes teoremine dayanmaktadırlar. Bayes teoremi rassal bir sürece bağlı olarak ortaya çıkan rasgele bir A olayı ile diğer bir rasgele B olayı için koşullu olasılıklar ve marjinal olasılıklar arasındaki ilişkiyi tanımlar. Matematiksel gösterimi A ve B olaylarının marjinal olasılıkları sırasıyla $P(A)$ ve $P(B)$ olmak üzere, $P(A/B)$ ve koşullu olasılıkları eşitlik 4.1.'de verilmiştir. (Stigler, 1982)

$$P(A/B) = \frac{P(B/A)P(A)}{P(B)}, \quad P(B) > 0 \quad (4.1.)$$

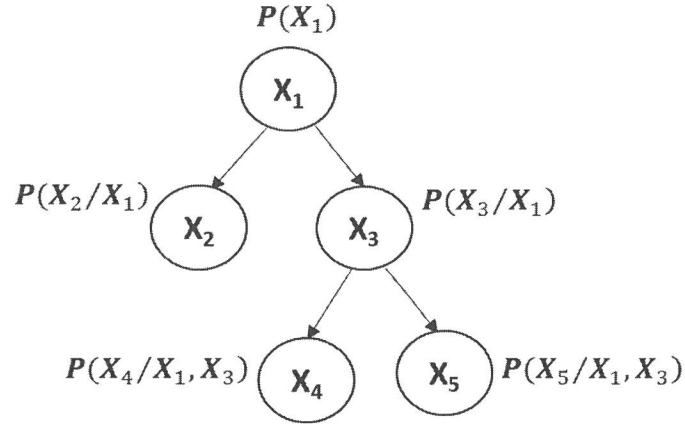
Burada, B olayı gerçekleştiğinde A olayının gerçekleşme olasılığı $P(A/B)$ ile A olayı gerçekleştiğinde sonra B olayının gerçekleşme olasılığını $P(B/A)$ ile temsil edilir.

4. 2.1 Bayes Ağ

Bayes ağları, koşullu olasılık dağılımını esas alan ağ modellerinin içerisinde yer almaktadır. Bayes ağları değişkenler arasındaki koşullu olasılık ilişkilerini ve ortak olasılık dağılımını bir ağ yapısı üzerinde göstermektedir. Bu ağ modelinde düğümler değişkenleri, düğümler arasındaki oklar ise değişkenler arasındaki bağımlılığı ifade eder. 5 düğümlü 3 katmanlı basit bir Bayes ağ modeli Şekil 4.1’de verilmiştir. Burada X_1 düğümü, X_2 ve X_3 değişkeninin ebeveyni, X_3 düğümü X_4 ve X_5 değişkeninin ebeveynidir. Bu gösterimde X_2 ve X_5 değişkenleri arasında okla gösterilen bir ilişki olmadığından değişkenler bağımsızdır denir. $P(X_1)$, X_1 düğümünün olasılığını, $P(X_2/X_1)$, X_1 biliniyorken X_2 ’nin olasılığını, $P(X_3/X_1)$, X_1 biliniyorken X_3 ’ün olasılığını, $P(X_4/X_1, X_3)$, X_1, X_3 biliniyorken X_4 ’ün olasılığını, $P(X_5/X_1, X_3)$, X_1, X_3 biliniyorken X_5 ’nin olasılığını, göstermektedir. Bayes ağındaki tüm düğümlerinin ortak olasılığı eşitlik 4.2.’deki formül kullanılarak hesaplanabilir. (Domingos vd., 1997)

$$P(X_1, X_2, \dots, X_k) = P(X_k/X_1, \dots, X_{k-1})P(X_{k-1}/X_1, \dots, X_{k-2})P(X_2/X_1)P(X_1) \quad (4.2.)$$
$$= \prod_{i=1}^k P(X_i/X_1, \dots, X_{i-1})$$

Burada, $X_1, X_2, \dots, X_k$ ağdaki tüm düğümleri göstermektedir. X_i düğümünün ebeveynleri $X_1, X_2, \dots, X_k$ düğümleriyle X_i düğümü arasında ilişkili olan tüm düğümleri içermelidir.



Şekil 4. 1. Bayes Ağ Modeli

Bayes Ağ sınıflandırma yöntemi iki parametreye ihtiyaç duymaktadır; Ağın koşullu olasılık dağılımlarını bulmak için bir tahmin edici ve ağın öğrenmesi için bir arama algoritması.

4.2.2. Naive Bayes

Naive Bayes sınıflandırıcısı, değişkenlerin bağımsızlığı varsayımı altında çalışan olasılıksal bir sınıflandırma yöntemidir. Değişkenler arasında önem sıralaması yapmadan tüm değişkenlerin aynı oranda öneme sahip olduğunu varsayar. Bu yöntemle sınıflandırmada en yüksek olasılığa sahip olan gözlemlerin oluşturduğu sınıf, hedef sınıf olarak tespit edilir (Lewis, 1998).

$$Y_{\text{hedef}} = \underset{y_j \in Y}{\operatorname{argmax}} P(y_j/x) \quad (4.3.)$$

Eşitlik 4.3.'de Y_{hedef} , hedef sınıfı, y olası durumu hesaplanacak değişken y_j değişkeni j . olası durumu ve x değişkeni de sınıfı tespit edilecek olan girdiyi ifade etmektedir. Girdi birden çok özelliğe sahip olduğu durumda Bayes formülü farklı bir forma dönüşür. Girdi verisinin sınıf tahmini tüm özelliklerin koşullu olasılıklarının çarpımından eşitlik 4.4. yardımıyla hesaplanmaktadır.

$$P(x_1, x_2, \dots, x_k / y_j) = \prod_{i=1}^k P(x_i / y_j) \quad (4.4)$$

Burada X girdisi için k tane özellik $X = x_1, x_2, \dots, x_k$ şeklinde ifade edilmektedir. Bayes teoremi gözlemlerin koşullu olasılık değerini bulurken Naive Bayes sınıflandırıcı hedef sınıfı bulmayı amaçlar. Bayes teoreminde paydada bulunan değer, tüm hedef durumların olasılık hesaplarında ortak olduğundan Naive Bayes sınıflandırmada ihmal edilir. Bu durumda birden fazla özelliğe sahip girdilerde hedef değişkeninin formülü eşitlik 4.5' deki gibi ifade edilir.

$$Y_{hedef} = \underset{y_j \in Y}{argmax} (P(y_j) \prod_{i=1}^k P(x_i / y_j)) \quad (4.5)$$

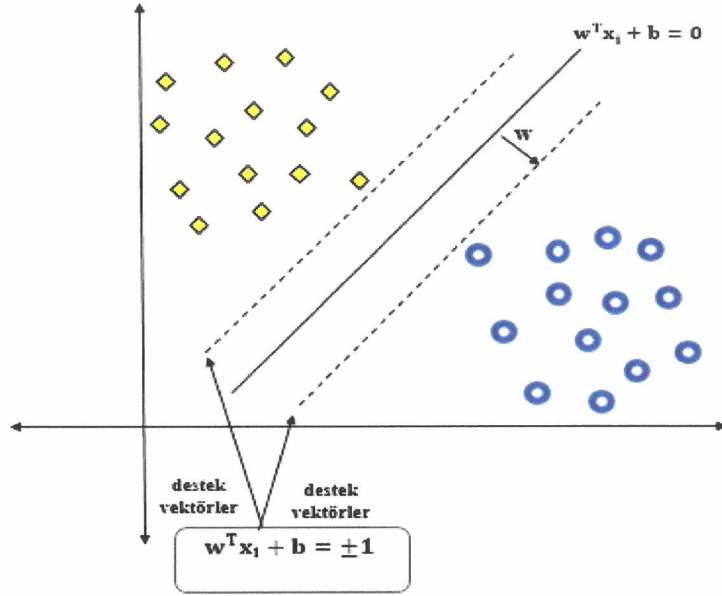
Naive Bayes sınıflandırma yönteminde veri kümesindeki değişkenlerin gerçek dağılımına modelin daha iyi uyum sağlayabilmesi için normal dağılım yerine çekirdek (kernel) tahmin edici kullanılmaktadır. Bu nedenle Naive Bayes sınıflandırıcının tek parametresi çekirdek tahmin edicisidir.

4.2.2 Destek Vektör Makinesi

Destek Vektör Makinası (DVM) sınıflandırma problemlerinde kullanılan, bir danışmanlı öğrenme yöntemidir. DVM yöntemi ile oluşturulan modellerde n boyutlu bir veri setini sınıflara ayırma işlemi sırasında sınıflar arasındaki uzaklıkların maksimum olması amaçlanır. Bu yolla DVM eğitim setindeki hataları minimize ederek sınıflandırma veya regresyon problemleri için destek vektörleri belirlemek için kullanılır (Vapnik, 1988). Bu destek vektörler belirli bir güven aralığında veri setini en iyi şekilde sınıflara ayırabilecek hiper düzlemleri oluşturmak için kullanılır. Veri setinin doğrusal olarak ayrılıp ayrılamamasına göre destek vektör makineleri iki şekilde ele alınır: doğrusal ve doğrusal olmayan DVM.

4.2.2.1 Doğrusal Destek Vektör Makinesi

Veri seti doğrusal olarak ayrılabilirdiği durumlarda Doğrusal DVM kullanılarak sınıflandırma yapılır. Doğrusal DVM sınıflandırıcıya bir örnek Şekil 4. 2'de verilmiştir.



Şekil 4. 2. Doğrusal DVM örneği

Şekilde iki farklı gruptaki veriyi en iyi şekilde ayırabilecek destek vektörler gösterilmiştir. Buradaki destek vektörler, aralarındaki uzaklık maksimum olacak şekilde ayrılır. Destek Vektörler 4.6. eşitsizliği ile verilen koşullar altında iki destek vektör arasındaki mesafeyi maksimum yapacak şekilde eşitlik 4.7. kullanılarak oluşturulur (Cortes ve Vapnik, 1995).

$$w^T x_i + b \begin{cases} \geq 1 \\ \leq -1 \end{cases} \quad \begin{cases} y = 1 \\ y = -1 \end{cases} \quad i = 1, \dots, n \quad (4.6.)$$

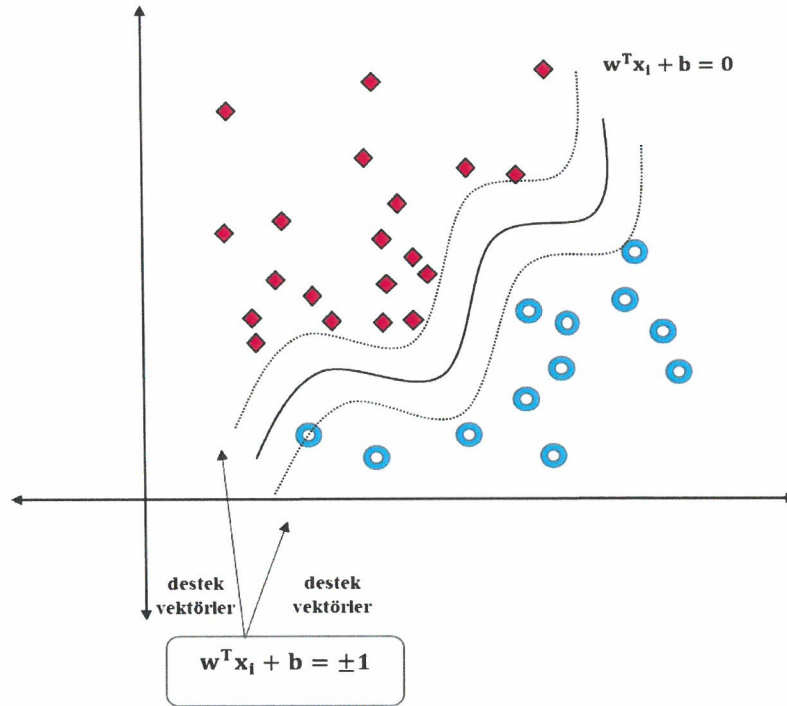
$$\max \frac{2}{\|w\|} \quad (4.7.)$$

Eş.6'da, n adet değişkene sahip x_i için girdi vektörünü; $y_i \in \{-1, +1\}$ sınıfları temsil eden çıktı vektörünü temsil etmektedir. Burada w , hiper düzlemin normali yani destek vektörlerin oluşturulmasında kullanılan ağırlık vektörünü; b , yanlılık katsayısını göstermektedir.

4.2.2.2. Doğrusal Olmayan Destek Vektör Makinesi

Verilerin doğrusal olarak ayrılamadığı durumlarda doğrusal sınıflandırıcılar yetersiz kalmaktadır ve doğrusal olmayan sınıflandırıcılara ihtiyaç duyulmaktadır. Doğrusal olmayan DVM sınıflandırma modeline ilişkin bir örnek Şekil 4. 3’de verilmiştir. Bu gibi durumlarda, doğrusal olarak ayrılmayan orjinal giriş uzayı doğrusal olmayan bir dönüşüm fonksiyonu yardımıyla daha yüksek boyutlu ve doğrusal olarak ayrılabilen bir özellik uzayına genişletilir. Çekirdek fonksiyonu yardımıyla bu dönüşüm yapılır. Bu çekirdek fonksiyonlarını kullanırken Mercer Teoremine dayanmaktadır. Mercer Teoremi yüksek boyutta çalışılırken çekirdek fonksiyonların her zaman iki girdi vektörünün iç çarpımı şeklinde ifade edilmesini sağlamaktadır (Tan vd. 2006). Bu çekirdek fonksiyon eşitlik 4. 8.’de verilen matematiksel formül ile hesaplanmaktadır.

$$\phi(x_i)^T \phi(x_j) = K(\vec{x}_i, \vec{x}_j) \quad (4.8.)$$



Şekil 4. 3. Doğrusal Olmayan DVM Örneği

Doğrusal olmayan DVM yönteminde çekirdek fonksiyonları yardımı ile veri seti n boyutlu uzaydan daha yüksek boyutlu (z boyutlu) uzaya taşınır. Sınıfları ayırma işleminde hiper düzlemi bulmak için kullanılan destek vektörler z boyutlu uzayda oluşturulur.

Burada $\phi(x)$ değeri daha yüksek boyutlu uzayda hesaplanabilir bir fonksiyon olmadığından destek vektör sınıflandırma için çekirdek fonksiyonları kullanılmaktadır. Çok boyutlu uzayda iç çarpımları veren uygun çekirdek fonksiyonu ile dönüşüm yaparak doğrusal olmayan DVM ile sınıflandırma işlemi yapılır.

İkili sınıflandırma modeline sahip Destek Vektör sınıflandırma yönteminde farklı çekirdek fonksiyonları kullanılmaktadır. Probleme göre doğru çekirdek fonksiyonunun belirlenmesi sınıflandırma başarısı açısından kritik bir öneme sahiptir. Hangi çekirdek fonksiyonun hangi durumda kullanılmasına dair kesin bir yargı bulunmamasıyla birlikte işlem kolaylığı açısından polinom fonksiyon tercih edilmektedir. Fakat tüm çekirdek fonksiyonları uygulanıp daha yüksek performans sağlayan çekirdek fonksiyonunun kullanılması daha doğru sonuçlara götürecektir. Polinom fonksiyon seçiminde polinom derecesinin belirlenmesi araştırmacı tarafından belirlenir. Fakat polinom derecesi arttıkça hesaplanabilme zorluğu olduğu da belirtilmektedir. Normalleştirilmiş polinom çekirdek fonksiyonu polinom çekirdek fonksiyonunun genelleştirilmiş hali olduğu bilinmektedir. Radyal tabanlı çekirdek fonksiyonlarda kullanılan parametre ise Gaus genişliğini ifade eden sigma parametresidir. Sigma parametresi için de en uygun değerin ne olduğu hakkında kesin bir yargı bulunmamaktadır. PUK kerneli uygulanırken matematiksel alt yapı anlamında daha karmaşık olmasına rağmen literatürde daha iyi sonuçlar verdiği bilinmektedir. (Kavzoğlu, Çölkesen, 2010) Yaygın kullanılan çekirdek fonksiyonları Çizelge 4.1’de gösterilmektedir.

Çizelge 4. 3. Çekirdek Fonksiyonları

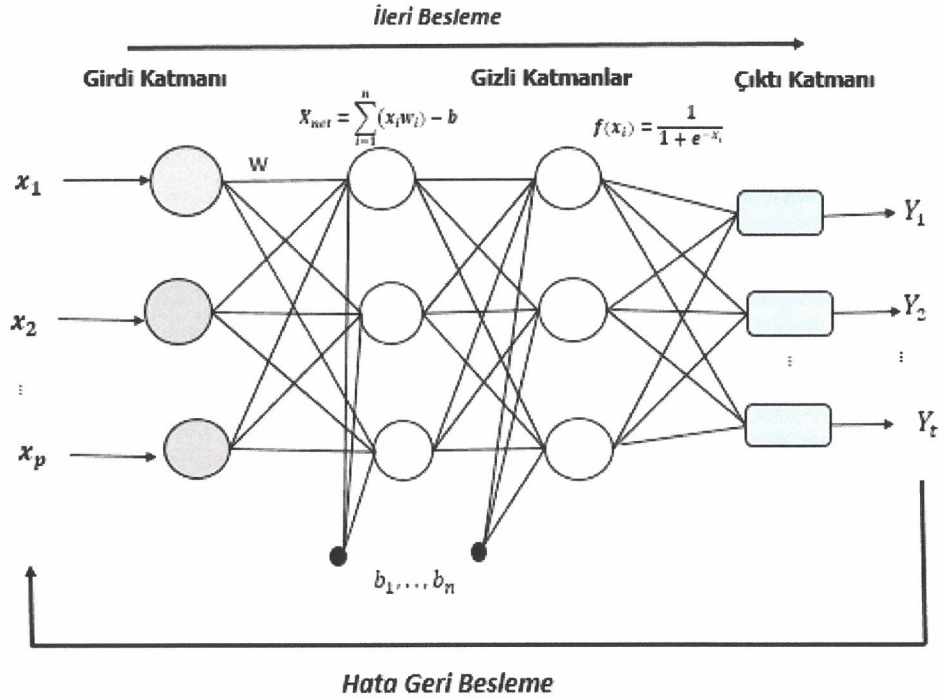
<i>Fonksiyon</i>	<i>Matematiksel İfade</i>
Polinom Fonksiyon	$K(x, y) = ((x \cdot y) + 1)^t$
Normalleştirilmiş Polinom Fonksiyon	$K(x, y) = \frac{((x \cdot y) + 1)^t}{\sqrt{((x \cdot x) + 1)^t + ((y \cdot y) + 1)^t}}$
Radyal Tabanlı Fonksiyon	$K(x, y) = e^{-\gamma (x - x_i) ^2}$
PUK Fonksiyon	$K(x, y) = \frac{1}{\left[1 + \left(\frac{2 \cdot \sqrt{\ x - y\ ^2 2^{(1/\omega)} - 1}}{\sigma} \right)^2 \right]^w}$

4.2.3 Çok Katmanlı Algılayıcılar

Yapay sinir ağları (YSA) insan öğrenmesini taklit etmek amacıyla beyindeki nöronlardan esinlenerek oluşturulmuş bir ağ modelidir. Ele alınan probleme göre farklı ağ mimarileri ve bu mimarilere özel eğitim algoritmaları mevcuttur. Bir sinir ağında eğitimin amacı, girdiyi en iyi ifade edecek modeli oluşturacak ağırlık vektörlerinin değerlerini tespit etmektir. Bu ağırlık vektörleri kullanılarak oluşturulan modelin başarı performansı test verileri yardımıyla ölçülmektedir. Model başarılıysa eğitim sonlandırılır. Modelin başarısı yeterli değilse eğitim tekrarlanır. Sınıflandırma için kullanılan en popüler eğitim algoritması Çok Katmanlı Algılayıcılarıdır (ÇKA). ÇKA ile eğitilen ağ modelinde bir girdi katmanı, bir veya birden fazla gizli katman ve bir de çıktı katmanı bulunmaktadır. İlk katman olan girdi katmanı girdi sinyallerini alır ve gerekli aktivasyon fonksiyonlarından geçirerek gizli katmana aktarır. Gizli katmanda işlenen sinyaller bir eşik fonksiyonu yardımıyla çıktı katmanına iletilir. Ağ modelini oluştururken, parametre güncellenmesi için alt örneklem sayısı, iterasyon sayısı, gizli katman sayısı, gizli katmanlardaki nöron sayıları, öğrenme oranı, aktivasyon fonksiyonu, eşik değeri ve durdurma kriteri gibi parametrelere ihtiyaç duyulur. Bu parametrelerin başlangıç değerleri deneme yanılma yoluyla veya rasgele belirlenebilir. Ağırlık vektörlerinin değerleri eğitim süresince

güncellenir ve önceden tanımlanan durdurma kriterine göre eğitimin sonlandırılır.

Şekil 4. 4’de ÇKA yapısı ağ yapısı için örnek bir model verilmiştir.



Şekil 4. 4. ÇKA Modeli

Burada $x_i; i = 1, \dots, p$ gözlemleri modelin p değişkenli girdi değerlerini, $h_k; k = 1, \dots, n$ ise n tane gizli katmanı, w_{ij} 'ler girdi katmandaki i . nöronu ve ilk gizli katmandaki j . nöron arasındaki bağlantıyı ifade eden ağırlık vektörlerine karşılık gelir. $b_k; k = 1, \dots, n$ ise yanlılığı (bias) değerlerini, $Y_t; t = 1, \dots, c$ ise çıktı katmanındaki nöron sayılarını ifade eder. (Haykin S., 2009). Modeldeki tüm girdi toplamı eşitlik 4.9. ile hesaplanmaktadır.

$$X_{net} = \sum_{i=1}^n (x_i w_i) - b \quad (4.9.)$$

Burada, X_{net} , bir toplam fonksiyonudur ve girdi katmandaki nöronların ağırlık vektörüyle çarpımlarından ve eşik değerinden elde edilir. Bu toplam değer bir aktivasyon fonksiyonundan geçirilir. En çok kullanılan aktivasyon fonksiyonu Eş.10'da verilen sigmoid fonksiyonudur (White, 1989).

$$f(x_i) = \frac{1}{1+e^{-x_i}} \quad (4.10.)$$

Toplam fonksiyonundan elde edilen net girdi değeri bir aktivasyon fonksiyonundan geçirilerek çıktı katmanına iletilir. Ele alınan probleme göre çıktı katmanında modelin çıktı değeri tahmin edilir. Gerçek çıktı ile modelin tahmin ettiği çıktı arasındaki fark sisteme hata olarak geri beslenir. Ağırlık vektörlerinin değerleri hatayı minimize edecek şekilde güncellenir ve aynı bu işlem hata değeri belirli bir eşik değerinin altına düşene kadar tekrarlanır (White, 1989). Modelin hatası eşitlik 4.11. yardımıyla hesaplanır.

$$\varepsilon_t = \frac{1}{2} \sum_{k=1} (g(t) - O_i(t))^2, \quad t = 1, \dots, n \quad (4.11.)$$

Burada ε_t , modelin hatasını $g(t)$, gerçek çıktı değeriyle $O_i(t)$, modelin çıktı değerini ifade etmektedir (Haykin, 2009). Bu süreç eğitim süreci olarak tanımlanır.

4.2.4 Karar Ağaçları

Karar Ağaçları (KA) kural tabanlı bir sınıflandırma yaklaşımıdır. Sınıflandırma ve regresyon problemlerinde sıklıkla kullanılır. Bir dizi kurallar çevresinde veri setini alt kümelerle bölmeyi hedefler.

KA yapısı düğümler, dallar ve yapraklar olmak üzere 3 temel bileşenden oluşur. Düğümler veri setindeki sınıflandırılacak niteliklerden meydana gelir, kök düğüm ve iç düğümler olarak iki alt grupta incelenir. Sınıflandırmaya kök düğüm ile başlanır. Veri setindeki en ayırt edici nitelik kök düğüm olarak belirlenir. Kök düğümün nasıl belirleneceğine karar vermek için belirsizliğin rastgeleliğinin ve beklenmeyen olası durumların bir ölçüsü olan entropi değerini kullanmaktadır.

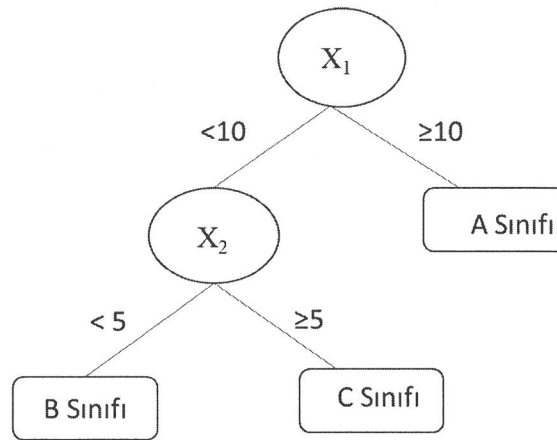
Entropi değerinin büyük olması düzensizliğin fazla olduğunu göstergesidir. Sınıflandırma problemlerinde öncelikle sınıf değişkeninin aldığı değerlere göre sistemin entropi değeri hesaplanır. Ardından sistemin entropi değeri üzerinde bilgi kazanımı en yüksek olan değişken belirlenir ve bu değişken kök düğüm olarak atanır.

Kök düğümün çıktıları yani bilgi kazancı en yüksek olan değişkenin değerleri iç düğümleri oluşturur. İç düğümlerin entropi değerleri kullanılarak benzer işlemler

tekrarlanır ve alt dallanmalar oluşturulur. Değişkenin aldığı her hangi bir değer için sınıf değişkeninin değeri sabit olduğunda entropi sıfır değeri alır ve dallanma sonlandırılır. Entropinin sıfır değeri alması belirsizliğin ortadan kalkması anlamına gelir ve KA'nın yaprakları oluşur. Yapraklar veri seti setindeki sınıf etiketlerini gösterir.

Değişken sayısı fazla ve değişkenlerin aldığı değerler sürekli olduğu durumlarda daha karmaşık KA modelleri ortaya çıkar. Karmaşık yapılarda birbirini tekrarlayan ve sınıflandırma doğruluğuna katkısı olmayan dallanmalar meydana gelebilir. Bu da aşırı öğrenmeye veya modelin karmaşıklığa neden olabilir. Bu problemi ortadan kaldırmak için KA yöntemleri budama işlemini kullanır. Budama, sınıflandırma doğruluğuna katkısı olmayan ve karmaşık yapıya neden olan dallanmaların karar ağacından atılma işlemidir. Bu sayede daha anlaşılabilir ve daha sade bir KA yapısı oluşturulur (Quinlan, 1986).

Şekil 4. 5' te X_1 , X_2 , girdi değişkenleri A,B,C de sınıfları temsil etmek üzere iki değişkenli 3 sınıflı basit bir KA modeline örnek verilmiştir.



Şekil 4. 5. Örnek Karar Ağacı Modeli

Şekil 4. 5'te X_1 , kök düğümü, X_2 iç düğümü, yapraklar ise A, B, C sınıflarını temsil etmektedir. X_1 kök düğüm olarak belirlendiği için sınıflandırmaya bu değişkenden başlanır ve değişkenin değerine göre dallanmalar oluşturulur. Bu örnekte X_1 'in aldığı değerleri 10'dan büyük eşit veya 10'dan küçük olma kuralına göre ikiye ayrıldığından iki dallanma meydana gelmiştir. X_1 değişkeninin değeri 10'dan büyük eşit olanlar A sınıfına atanmıştır. Değeri 10'dan küçük olanlar için X_2 değişkenin entropi değeri üzerinden

değişken değerlerinin bilgi kazanç miktarı hesaplanır. Çok değişkenli veri setlerinde her iç düğüm için benzer işlemler değişken değerlerinin entropileri sıfır değerini alana kadar tekrarlanır. Bu örnekte X_2 değişkeninin değeri 5'den büyük eşit olanların tamamı B sınıfında yer aldıkları için entropi değeri sıfır bulunmuş ve yaprak oluşturulmuştur. Benzer şekilde değeri 5'den büyük olanların tamamı da C sınıfında yer aldıkları için C sınıfı yaprak olarak belirlenmiştir.

Literatürde KA yöntemlerinin temel mantığını kullanan ancak kendi içinde farklılaşmış bazı algoritmalar mevcuttur. Bu çalışmada galaksi sınıflandırması amacıyla temel karar ağacı yöntemlerinden J48, LMT, Rasgele Orman algoritmaları ele alınmıştır.

4.2.4.1 J48 Algoritması

Sınıflandırma amacıyla kullanılan KA yöntemlerinin temel algoritması C4.5 algoritmasıdır. Diğer algoritmalar C4.5 üzerinden geliştirilmiştir. J48 ise C4.5 algoritmasının Weka platformundaki adıdır. Amacı en yüksek bilgi kazanımı sağlayan tahmin edici değişkeni belirleyerek sınıfları tespit etmektir. Bilgi kazanımı en yüksek olan değişken değerini bulurken eşitlik 4.12. yardımıyla hesaplanan entropi değerini kullanılır. Entropi değeri belirlendikten sonra her bir sınıf için değişkenlerin bilgi değeri ve ardından her bir değişkenin bilgi kazanımını eşitlik 4.13. yardımıyla hesaplanır. Eşitlik 4.13'de değişkenlerin bilgi değerleri ve Eşitlik 4.14.'te bilgi kazanım değerinin belirlendiği formül verilmiştir (Quinlan, 1986).

$$Entropi = - \sum_{i=1}^m (p_i \log_2 p_i) \quad (4.12.)$$

$$Bilgi_A(D) = - \sum_{j=1}^v \frac{|D_j|}{|D|} \times Entropi \quad (4.13.)$$

$$Kazanç(A) = Bilgi_A(D) - Entropi \quad (4.14.)$$

Burada m, hedef değişkenin alabileceği değerlerin sayısını; A, tahmin edici değişkeni; D, hedef sınıfı; v, tahmin edici değişkenin alabileceği değerlerin sayısı; p_i , i. sınıfın olasılık değerini göstermektedir.

Weka programında J48 algoritması uygulanırken budama parametresi, ikili veya çoklu ayrışım türü, yaprak başına düşen minimum örnek sayısı, budama için güven faktörü değerinin belirlenmesi gerekmektedir. Güven faktörü (confidence factor) budama kriteri olarak kullanılmaktadır. Bu değer budama için eşik değerini ifade eder.

4.2.4.2 Lojistik Model Ağacı

Lojistik Model Ağacı (LMT), lojistik regresyon ve karar ağacı yöntemlerinin birleşimi ile elde edilen danışmanlı makine öğrenmesi yöntemidir. LMT temel olarak, yapraklarda lojistik regresyon fonksiyonlarına sahip bir karar ağacı yapısından oluşur. Dalların J48 Algoritmasının çalışma prensibine göre gerçekleştirilir.

LMT, veri setindeki tüm özelliklerin kapsayan S örnek uzayını eşitlik 4.15.'de verilen kurallara ışığında S_t adet ayrık alt bölgeye böler. Her bölge ağaçtaki bir yaprak ile temsil edilir.

$$S = \bigcup_{t \in T} S_t, \quad S_t \cap S_{t'} = \emptyset \quad \text{for } t \neq t' \quad (4.15.)$$

Burada S , verinin tüm özelliklerini içeren örnek uzay; T , ağacın yapraklarının kümesini t ve t' bu kümenin elemanları olan farklı yaprakları temsil eder. Lojistik regresyon fonksiyonları ile temsil edilen sınıflar eşitlik 4.16 yardımı ile belirlenir.

$$f(x) = \sum_{t \in T} f_t(x) \cdot I(x \in S_t) \quad (4.16.)$$

Burada $I(x \in S_t)$, 0 ve 1 değerlerini alan bir fonksiyondur. Eğer x girdisi, S_t alt kümesinin elemanı ise fonksiyon 1, değilse 0 değerini üretmektedir. $f_t(x)$ ise yaprakları temsil eden lojistik regresyon fonksiyonudur (Landwehr vd, 2005)

LMT sınıflandırma yöntemi, bir düğümün bölünebileceği minimum örnek sayısı, ikili veya çoklu ayrışım türü, logit boost durdurma kriteri için hata belirleme, logitboost için ağırlık düzeltme kriteri vb. parametrelere ihtiyaç duymaktadır.

Logitboost algoritması aşırı uyumu engellemek için kullanılır. Küçük değerler alması daha yüksek güvenilirlikte sınıflandırma yapmayı sağlar. Genelde 0.1 ve 0.01 değerlerini alır.

4.2.4.3 Rasgele Orman

Rasgele orman (RO) algoritmasının çalışma mantığı J48 algoritmasına dayanmaktadır. J48 den farklı olarak tek tek oluşturulan karar ağaçları bir araya gelerek karar ormanı oluşturur. Bu sınıflandırma yönteminde amaç birden fazla karar ağacı kullanılarak her biri farklı eğitim kümelerinde eğitilmiş olan çok sayıda ağacın kararlarını birleştirmektir. Bu kararlar çoğunluk oyuyla birleştirilir. Bu sayede sınıflandırma performansının yükseltilmesi hedeflenmektedir. Algoritmayı başlatmak için her bir düğümde kullanılan değişkenlerin sayısı ve geliştirilecek ağaçların sayısı kullanıcı tarafından belirlenmektedir. Bu yöntemde dallanma kriterlerinin belirlenmesinde en yaygın kullanılan teknik Gini indeksidir

RO ile ağaç üretmek için kullanılan Gini indeksi eşitlik 4.17.'de verilmiştir.

$$\sum \sum_{i \neq j} (f(k_i, T)/|T|)(f(k_j, T)/|T|) \quad (4.17.)$$

Bu eşitlikte, T ; eğitim veri setini k_i ; rastgele seçilen gözlem değerinin sınıfı ve $(f(k_i, T)/|T|)$, seçilen örneğin k_i sınıfına ait olma olasılığını göstermektedir (Breiman, 2001).

4.3 Performans Değerlendirme Kriterleri

İstatistiksel öğrenme ve makine öğrenmesine dayalı sınıflandırma yöntemlerinin sınıflandırma performanslarını değerlendirmek için Sınıflandırma Başarı Yüzdesi, Kappa İstatistiği ve karmaşıklık matrisinden elde edilen performans kriterleri kullanılmaktadır.

Karmaşıklık matrisinden elde edilen performans kriterleri sınıfların ne kadar doğru sınıflandırıldığını değerlendirirken, Başarı Yüzdesi ve Kappa İstatistiği modelin doğruluğu hakkında bilgi vermektedir.

4.3.1 Sınıflandırma Başarı Yüzdesi

Sınıflandırma Başarı Yüzdesi (BY), sınıflandırma yöntemleriyle geliştirilen modelin doğruluğunun bir ölçüsüdür. Modelde doğru sınıflandırılan gözlemlerin tüm gözlemlere oranından elde edilir.

$$SBY = \frac{G_d}{G_T} \times 100 \quad (4.18.)$$

G_d , Doğru sınıflandırılan gözlemlerin sayısı, G_T : Tüm gözlemlerin sayısı olmak üzere sınıflandırma başarı yüzdesinin formülü eşitlik 4.18.'de görülmektedir.

4.3.2 Kappa İstatistiği

Kappa istatistiği (Kİ), gerçek gözlemler ile tahmin edilen gözlemler arasındaki uyumu ölçmek için kullanılan performans kriteridir. Kİ değerinin 1 olması model tahmini ile gerçek sınıfının tam olarak uyumlu olduğunu gösterir. Kİ değeri 1-0,8 aralığında çok iyi, 0,8-0,6 aralığında iyi, 0,6-0,4 aralığında orta, 0,4-0,2 aralığında zayıf düzeyde, 0,2-0 aralığında önemsiz düzeyde uyumu göstermektedir. Bu değerin 0'dan küçük olması ise uyumun rastgele olabileceği anlamına gelmektedir.

Kİ eşitlik 4.19. yardımı ile hesaplanmaktadır. Burada P_T , gözlemlenen uyumların toplam olasılığını; P_e , gözlemlerin tesadüfi uyumunun olasılığını; K ise Kappa istatistiğini göstermektedir (Wampold, 1989).

$$K = \frac{P_T - P_e}{1 - P_e} \quad (4.19.)$$

4.3.3 Karmaşıklık Matrisi

Sınıflandırma yöntemlerinin performansını ölçmek için kullanılan bir diğer yaklaşım da karmaşıklık matrisi (confusion matrix)'dir. Karmaşıklık matrisi (KM), tahmin edilen sınıflarla gerçek sınıflar arasındaki uyumu karşılaştırma imkânı sağlayan bir matris formudur. Sınıf düzeyleri üzerinden sınıflandırma sonuçları kullanılarak oluşturulur. Çizelge 4. 2'de ikili sınıflandırma sonuçlarına göre oluşturulan bir karmaşıklık matrisine

örnek verilmiştir. Burada gerçek sınıflar Doğru ve Yanlış olarak iki değer alırken, sınıflandırma sonucunda tahmin edilen değerlerde yine Doğru ve Yanlış şeklinde tahmin edilmiştir. Bu durumda tahmin edilen sınıf ile gerçek sınıf arasındaki ilişki 4 farklı duruma göre incelenir (Godbole, 2002).

Çizelge 4. 4. Karmaşıklık Matrisi

Karmaşıklık matrisi		Gerçek Sınıf	
		Doğru	Yanlış
Tahmin Sınıf	Doğru	TP	FP
	Yanlış	FN	TN

Burada TP; Gerçek sınıfı doğru olan gözlemlerin doğru tahmin edilmesi durumu, FP; Gerçek sınıfı yanlış olan gözlemlerin doğru tahmin edilmesi durumu, FN; Gerçek sınıfı doğru olan gözlemlerin yanlış tahmin edilmesi durumu ve TN; Gerçek sınıfı yanlış olan gözlemlerin yanlış tahmin edilmesi durumunu temsil etmektedir.

TP ve TN sınıflandırma yöntemleriyle oluşturulan modelin başarılı tahmin yaptığı durumlardır. Yani model gerçek sınıfı doğru olan gözlemleri doğru tahmin etmiş, gerçek sınıfı yanlış olan gözlemleri de yanlış olarak tahmin etmiştir. FN ve TN ise modelin hatalı tahmin yaptığı durumlara karşılık gelir. Bu durumlar istatistiksel testlerdeki 1. tip ve 2. tip hataya karşılık gelir. KM üzerinden hesaplanan Doğruluk (Accuracy), Hassasiyet (Precision), Duyarlılık (Recall) gibi performans kriterleri sınıflandırma yöntemlerinin karşılaştırılmasında sıklıkla kullanılan kriterlerdir.

Doğru sınıflandırılmış gözlemlerin tüm gözlemlere oranından eşitlik 4.20. yardımı ile hesaplanan *Doğruluk*, oluşturulan modelin tahminlerde ne kadar başarılı olduğunun bir ölçüsüdür. 1'e yakın değerler alması beklenir.

$$\text{Doğruluk} = \frac{TP+TN}{TP+FP+FN+TN} \quad (4.20.)$$

Modelin doğru olarak tahmin ettiği gözlemlerin gerçekte ne kadarının doğru olduğunu açıklayan *Hassasiyet* eşitlik 4.21. kullanılarak hesaplanır. Modelin pozitif sınıf etiketini tahmin etmekteki etkililiğinin bir ölçüsüdür.

$$\text{Hassasiyet} = \frac{TP}{TP+FP} \quad (4.21.)$$

Doğru sınıflandırılmış örnek sayısının, gerçekte doğru olan gözlem sayısına oranı olan *Duyarlılık* eşitlik 4.22. kullanılarak hesaplanır. Pozitif sınıfa ait gözlemlerin ne kadarının doğru tahmin edildiğinin bir ölçüsüdür.

$$\text{Duyarlılık} = \frac{TP}{TP+FN} \quad (4.22.)$$

4.4. Friedman Testi

İkiden fazla bağımlı grubu karşılaştırmak ve grupların arasında anlamlı farklılık olup olmadığını belirlemek için Friedman testi kullanılmaktadır. Tekrarlı ölçümlerde kullanılan iki yönlü varyans analizinin parametrik olmayan karşılığıdır. Friedman testinin H_0 hipotezi gruplar arasında anlamlı bir farklılık yoktur, H_1 alternatif hipotez ise gruplar arasındaki anlamlı bir farklılığın var olduğunu savunan hipotezlerdir. Friedman test istatistiği gözlemlere atanan sıra numaraları üzerinden hesaplanır. Sıralama işlemi her bloktaki gözlem değerlerine küçükten büyüğe doğru 1’den k (k, blok sayısı) ya kadar sıra numarası verilmesi şeklinde gerçekleştirilir. Elde edilen sıra toplamaları kullanılarak eşitlik 4.23.’de verilen formüle göre test istatistiği hesaplanır. Friedman testinin test istatistiği χ_R^2 , ve χ_{k-1}^2 n işlem sayısı, k blok sayısı olmak üzere k-1 serbestlik dereceli tablo değerini göstermektedir. $\chi_R^2 \geq \chi_{k-1}^2$ olduğunda H_0 reddedilir. Bu durumda χ_R^2 değerinin tesadüfe bağlı olarak ortaya çıkma olasılığı olan p değeri α anlamlılık düzeyinden de küçüktür sonucuna varılır. (Canküyer, Aşan 2005)

$$\chi_R^2 = \left[\frac{12}{nk(n+1)} \sum_{j=1}^k R_j^2 \right] - [3n(k+1)] \quad (4.23.)$$

Burada, n; işlem sayısını, k; blok sayısını ve R_j ; her bir grubun sıra toplamalarını ifade etmektedir (Friedman, 1937). Bu çalışma da Shapley konsanstrasyon bölgesindeki galaksileri sınıflandırmak için uygulanan yöntemler arasında anlamlı farklılığı incelemek için Friedman testi kullanılmıştır. Shapley bölgesine ait doğal gruplarının bulunduğu Abell grubu karşılaştırma için kontrol grubu olarak tanımlanmıştır. Yöntemler arasında anlamlı farklılıklar bulunduğu takdirde çoklu karşılaştırma testi yapılarak hangi gruplar arasında farklılıklar olduğu tespit edilir.

5. UYGULAMA

Bu çalışmada gökyüzünde 193.03 ile 216.03 boylam koordinatında -27.5 ile -37.65 enlem koordinatında yer alan Shapley Konsantrasyon bölgesindeki 4215 galaksi, boylam, enlem, parlaklık, hız ve hızdaki sapma değişkenleri dikkate alınarak incelenmiştir. Shapley Konsantrasyon bölgesine ait galaksilerin doğal grupları George O. Abell tarafından kataloglanan Abell katalogunda yer almaktadır. Shapley Konsantrasyon bölgesinde ki galaksi gözlemlerine ilişkin veri seti Penn State Üniversitesi Astroistatistik merkezinden ([http 1](http://1)), Abell Kataloğuna ilişkin veriler ise Vizier veri tabanından ([http 2](http://2)) alınmıştır. Shapley Konsantrasyon bölgesindeki galaksilerin doğal gruplarını hangi Abell grubunda yer aldığını tespit etmek için IDL programı kullanılmıştır.

5.1. IDL Programlama ile Galaksilerin Doğal Gruplarının Tespiti

IDL programlama, sayısal verileri görselleştirmek, iki veri seti arasında karşılaştırma yapabilmek ve görsel verilerin işlenerek sayısal sonuçlar çıkarmak için kullanılan bir yazılım paketidir. Astronomik verilerin analizinde sıklıkla kullanılmaktadır.

Shapley Konsantrasyon bölgesindeki galaksiler tüm gökyüzünde 13 derecelik açıda yayılmıştır. Shapley Konsantrasyon bölgesindeki galaksilerin Abell kataloğundaki galaksi grupları ile karşılaştırılmasında IDL programı kullanılmıştır. IDL programlama da karşılaştırma yaparken yalnızca enlem ve boylam değişkenleri kullanılmaktadır. IDL programlama ile Shapley bölgesine galaksilerin A1633, A1648, A1736, A1771, A1802, A1816, A1846, A1857 olmak üzere Abell kataloğunda 8 farklı grupta sınıflandırıldığı tespit edilmiştir. Bu Abell gruplarının aralarındaki mesafeler Öklid uzaklıkları kullanılarak hesaplanmıştır. Aralarındaki Öklid uzaklığının daha küçük olması galaksilerin birbirine daha yakın olduğunu ifade eder (Gower J.C., 1982).

Enlem ve boylam değişkenleri dikkate alınarak hesaplanan Öklid uzaklıkları Çizelge 5.1'de verilmiştir.

Çizelge 5.1 Gruplar Arasındaki Öklid Uzaklıkları

	1633	1648	1736	1771	1802	1816	1846	1857
1633	0,00	2,27	9,14	11,83	15,02	18,52	20,38	21,89
1648	2,27	0,00	7,17	10,33	13,20	16,81	18,78	20,39
1736	9,14	7,17	0,00	4,24	6,09	9,79	11,92	13,67
1771	11,83	10,33	4,24	0,00	4,02	6,86	8,56	10,08
1802	15,02	13,20	6,09	4,02	0,00	3,75	6,02	7,90
1816	18,52	16,81	9,79	6,86	3,75	0,00	2,38	4,32
1846	20,38	18,78	11,92	8,56	6,02	2,38	0,00	1,94
1857	21,89	20,39	13,67	10,08	7,90	4,32	1,94	0,00

5.2. Weka Sınıflandırma Sonuçları

Doğal grubu tespit edilen galaksiler Weka programında istatistiksel ve makine öğrenme yöntemleri kullanılarak sınıflandırılmıştır. Bayes Sınıflandırıcılardan Bayes Ağ ve Naive Bayes, Karar Ağaçlarından J48, LMT, RO yöntemleri, DVM Sınıflandırıcı ve Yapay Sinir Ağları yöntemlerinden ÇKA kullanılmıştır. Weka programı veri setini arff. Dosyası olarak kabul etmektedir. Arff formatına dönüştürülen veri setinin bir parçası Çizelge 5. 2’de sunulmuştur. Gruplara göre galaksi sayılarının dağılım grafiği ise Şekil 5.1’de verilmiştir.

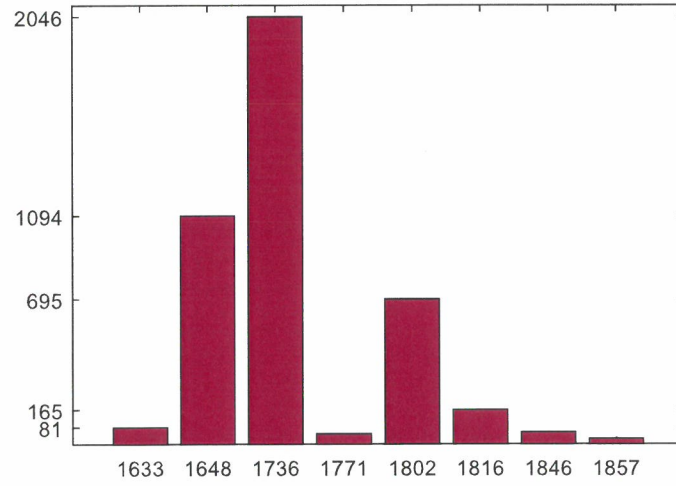
Weka da yöntemlerin uygulanma aşamasında sınıflandırma modellerini değerlendirmek için k-kat çapraz doğrulama testi (k-Fold Cross-validation) kullanılmıştır. Bu yöntem veriyi eğitim seti ve test yani doğrulama seti olarak k sayıda parçaya ayırır. Ayrılan her bir parça hem eğitim hem test için kullanılır. Bu çalışmada, k =10 olarak alınmış ve veri seti 10 parçaya ayrılmıştır. Her parça bir kez test seti ve geriye kalan 9 parça eğitim seti olarak kullanılmıştır. Yöntemlerin performansları doğrulama setine göre değerlendirilmiştir. Bu performans değerleri galaksilerin Abell kataloğundaki gruplara uygunluklarına göre değerlendirilmiştir.

Şekil 5.1’de verilen sınıfların dağılımı grafiğinde 1736 numaralı Abell grubunda 2046 galaksi bulunurken 1857 numaralı grupta 26 tane galaksinin yer aldığı görülmektedir. Abell gruplarından A1771 grubu içerisinde eliptik ve spiral galaksilerin, A1736 grubunda yalnızca spiral galaksilerin, A1802 grubunda hem eliptik hem de spiral galaksilerin, Abell 1648 grubunda ise yalnızca spiral galaksilerin olduğu bulunmaktadır (Struble ve Rood,

1987). Diğer gruplardaki galaksi biçimleri hakkında bilgiye ulaşılammıştır.

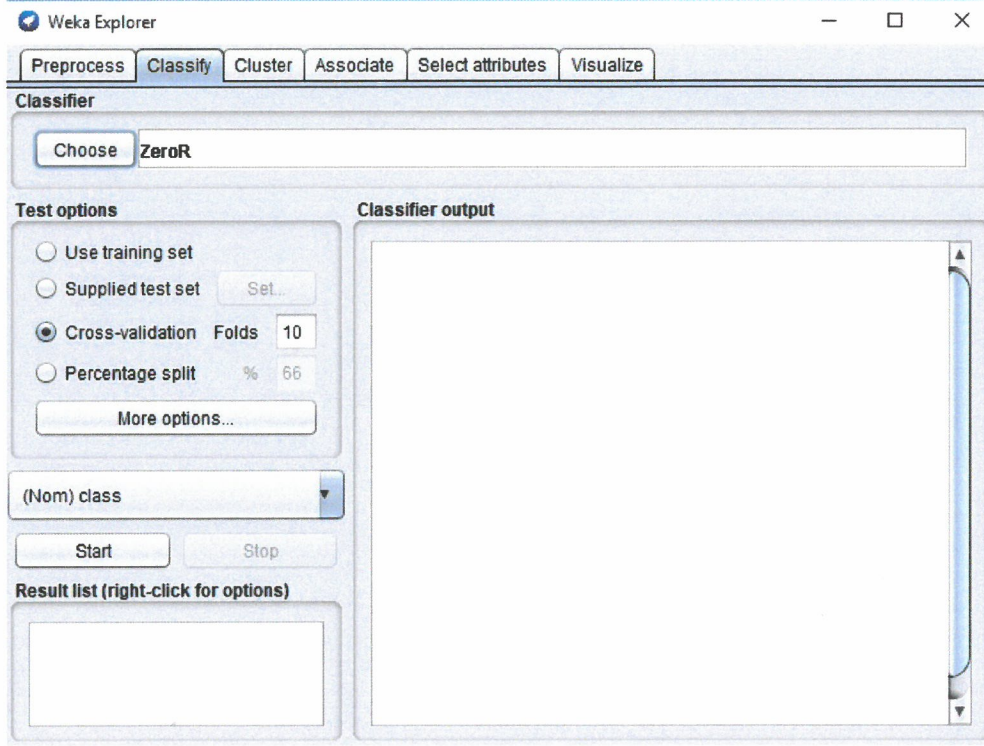
Çizelge 5. 2. Weka formatına dönüştürülmüş veri seti

```
@relation data
@attribute R.A numeric
@attribute DEC numeric
@attribute MAG numeric
@attribute V numeric
@attribute SIGV numeric
@attribute class {1633,1648,1736,1771,1802,1816,1846,1857}
@data
193.03,-32.85,15.23,15056,81,1648
193.04,-28.54,17.22,16995,32,1633
193.04,-28.23,17.29,21211,81,1633
193.05,-28.34,18.2,29812,37,1633
193.06,-29.84,12.55,2930,38,1633
193.06,-30.33,17.62,16042,104,1633
193.06,-30.12,16.87,17264,36,1633
193.08,-27.89,16,16975,57,1633
193.08,-30.51,17.9,31572,68,1633
193.13,-28.39,17.71,14806,36,1633
193.13,-28.95,15.8,16559,118,1633
193.14,-28.45,17.9,17032,56,1633
193.14,-30.48,18.89,16618,15,1633
193.14,-30.25,17.58,27457,15,1633
193.15,-31.27,13.9,16404,157,1648
193.15,-31.89,15.18,3543,138,1648
```



Şekil 5. 1. Abell gruplarında bulunan galaksi gözlem sayıları

Şekil 5. 2’de verilen görselde Weka platformunda sınıflandırma için kullanılan ara yüz gösterilmiştir.



Şekil 5. 2.Weka platformunda sınıflandırma aracı

5.2.1 Bayes Ağ

Weka platformunda Bayes ağının koşullu olasılık dağılımlarını tahmin etmek için Basit tahmin edici (Simple estimator) ve ağın öğrenmesi için arama algoritmasından Benzetimli Tavlama algoritması (Simulated Annealing) kullanılmıştır.

Sonuçlar incelendiğinde galaksilerin sınıflandırma başarı yüzdesi %95.27'dir. Bayes ağ için Kİ değeri 0.92'dir. Karmaşıklık matrisinden elde edilen performans kriterlerinden Doğruluk, Duyarlılık ve Hassasiyet değerleri sırasıyla 0.98, 0.76 ve 0.84'dür. Çizelge 5. 3'de verilen karmaşıklık matrisinde gerçek sınıflar ile tahmin edilen sınıflar arasında her bir sınıf için doğru ve yanlış sınıflandırılan gözlemler verilmiştir. Örneğin, gerçekte Abell 1736 sınıfında yer alan için 2046 gözlem varken Bayes Ağ ile bu gözlemlerin 2024 tanesi doğru tahmin edilmiştir.

Geriye kalan gözlemlerin 11 tanesi Abell 1633 sınıfında, 3 tanesi Abell 1771 grubunda ve 8 tanesi de Abell 1802 sınıfında olmak üzere yanlış tahmin edilmiştir.

Çizelge 5. 3. Bayes Ağ Yöntemi için Karmaşıklık Matrisi

Gerçek sınıf	Tahmin Sınıf							
	A1633	A1648	A1736	A1771	A1802	A1816	A1846	A1857
A1633	53	31	0	0	0	0	0	0
A1648	4	1077	11	2	0	0	0	0
A1736	0	11	2024	3	8	0	0	0
A1771	0	0	21	22	7	0	0	0
A1802	0	0	8	3	651	32	1	0
A1816	0	0	0	0	20	136	9	0
A1846	0	0	0	0	1	14	38	6
A1857	0	0	0	0	0	1	7	18

5.2.2. Naive Bayes

Naive Bayes sınıflandırmada çekirdek tahmin edici (kernelEstimator) kullanılmıştır. Yöntemin galaksileri sınıflandırmadaki başarısı %93.28 bulunurken Kİ değeri 0.92 elde edilmiştir. Naive Bayes için Doğruluk değeri 0.89, Duyarlılık değeri 0.77 ve Hassasiyet değerleri 0.84'dür. Naive Bayes sınıflandırma sonuçları Çizelge 5.4'te verilmiştir.

Çizelge 5. 4. Naive Bayes Yöntemi için Karmaşıklık Matrisi

Gerçek sınıf	Tahmin sınıf							
	A1633	A1648	A1736	A1771	A1802	A1816	A1846	A1857
A1633	58	23	0	0	0	0	0	0
A1648	53	1026	15	0	0	0	0	0
A1736	0	20	2009	1	16	0	0	0
A1771	0	0	17	23	10	0	0	0
A1802	0	0	14	0	626	55	0	0
A1816	0	0	0	0	33	129	3	0
A1846	0	0	0	0	1	13	43	1
A1857	0	0	0	0	0	0	8	18

5.2.3 Destek Vektör Makinesi

Weka platformunda uygun olan çekirdek fonksiyonlarını tespit edebilmek için tüm çekirdek fonksiyonları çalıştırılmıştır. Sonuçları Çizelge 5.5'te verilmiştir.

Çizelge 5. 5. Çekirdek Fonksiyonlarının Uygulama Sonuçları

Fonksiyon	Kappa	Uygulama Sonucu Modelin Başarı Yüzdesi
Polinom Fonksiyon	0.883	92.36 %
Normalleştirilmiş Polinom Fonksiyon	0.806	87.3072 %
Radyal Tabanlı Fonksiyon	0.510	71.4353 %
PUK Fonksiyon	0.910	94.0925

Bu sonuçlara göre en uygun çekirdek fonksiyonunun PUK çekirdek fonksiyonu olduğu görülmüştür. DVM sınıflandırma PUK çekirdek fonksiyonu kullanılmıştır.

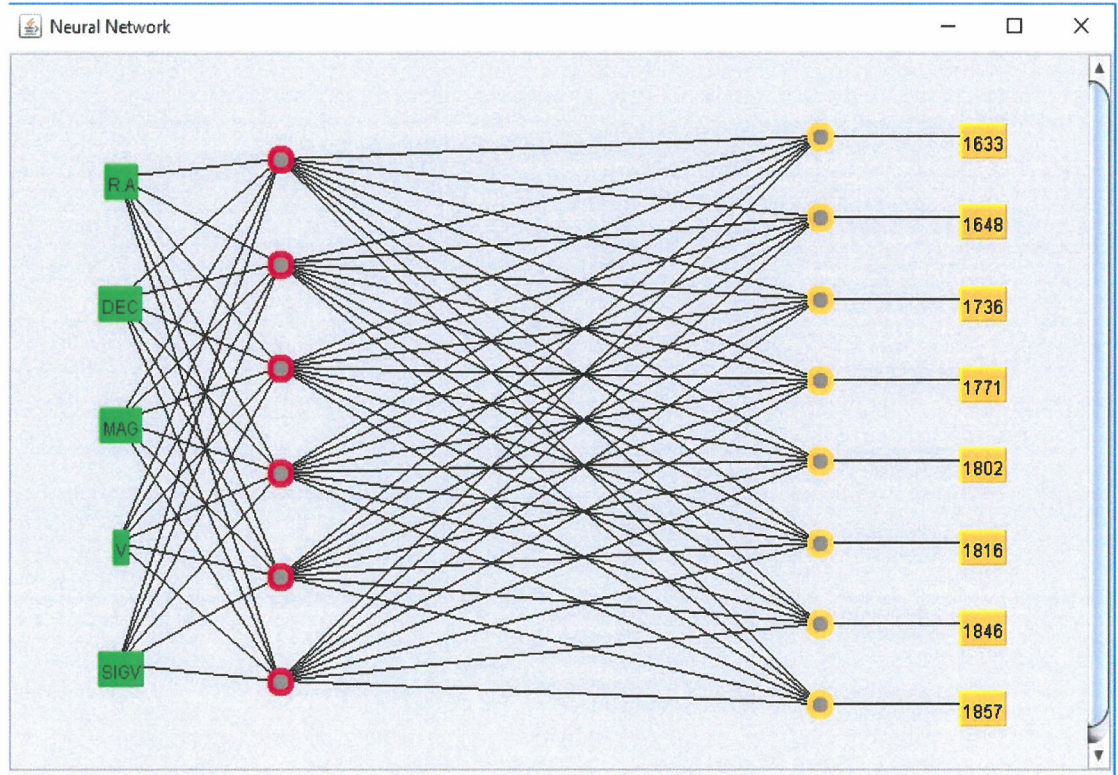
DVM yönteminin sınıflandırma başarı oranı %94.09'dur. Kİ ise 0,91'dir. Doğruluk 0.99, Duyarlılık 0.60 ve Hassasiyet 0,69 olarak bulunmuştur. Çizelge 5.6'da DVM yöntemi için tahmin sonuçları verilmiştir. DVM yönteminin Abell 1633 grubundaki galaksilerin hiçbirini doğru tahmin edemediği bu gruptaki tüm gözlemleri Abell 1648 grubunda olacak şekilde yanlış tahmin ettiği görülmüştür.

Çizelge 5. 6. Destek Vektör Sınıflandırma Yöntemi için Karmaşıklık Matrisi

Gerçek sınıf	Tahmin sınıf							
	A1633	A1648	A1736	A1771	A1802	A1816	A1846	A1857
A1633	0	81	0	0	0	0	0	0
A1648	0	1079	15	0	0	0	0	0
A1736	0	14	2031	0	0	0	0	0
A1771	0	0	29	5	16	0	0	0
A1802	0	0	24	0	668	3	0	0
A1816	0	0	0	0	31	133	1	0
A1846	0	0	0	0	1	12	45	0
A1857	0	0	0	0	0	3	18	5

5.2.4 ÇKA

Weka programında ÇKA sınıflandırma yöntemi uygulanırken gizli katman sayısı Wilcard kriterine göre belirlenmiştir (Morariu, 2017). Eğitim parametrelerinin optimal değerleri alt örneklem sayısı (batch size) 50, iterasyon sayısı 1000, durdurma kriterlerinden biri olan doğrulama eşiği 20, öğrenme oranı 0.3 olarak belirlenmiştir. Eğitim için oluşturulan çok katmanlı sinir ağı yapısı Şekil 5.3’de verilmiştir. Bu ağda kullanılan R.A, DEC, Mag, V, SIGV ağı girdi katmanındaki nöronları göstermektedir. Bu nöronlar Shapley bölgesine ait galaksilerin değişkenlerine (boylam, enlem, parlaklık, hız ve hızdaki sapma değerleri) karşılık gelmektedir. Çıktı katmanında ise Abell sınıfları (A1633, A1648, A1736, A1771, A1802, A1846, A1857) yer almaktadır.



Şekil 5. 3 Shapley konsanstrasyon bölgesine ait çok katmanlı sinir ağı yapısı

ÇKA sınıflandırma yönteminin Shapley bölgesindeki galaksileri sınıflandırma başarı yüzdesi % 96.46’dır. Kİ değeri 0.94’dür. Doğruluk değeri 0.99, Duyarlılık değeri 0.71 ve Hassasiyet değeri ise 0.92 olarak bulunmuştur.

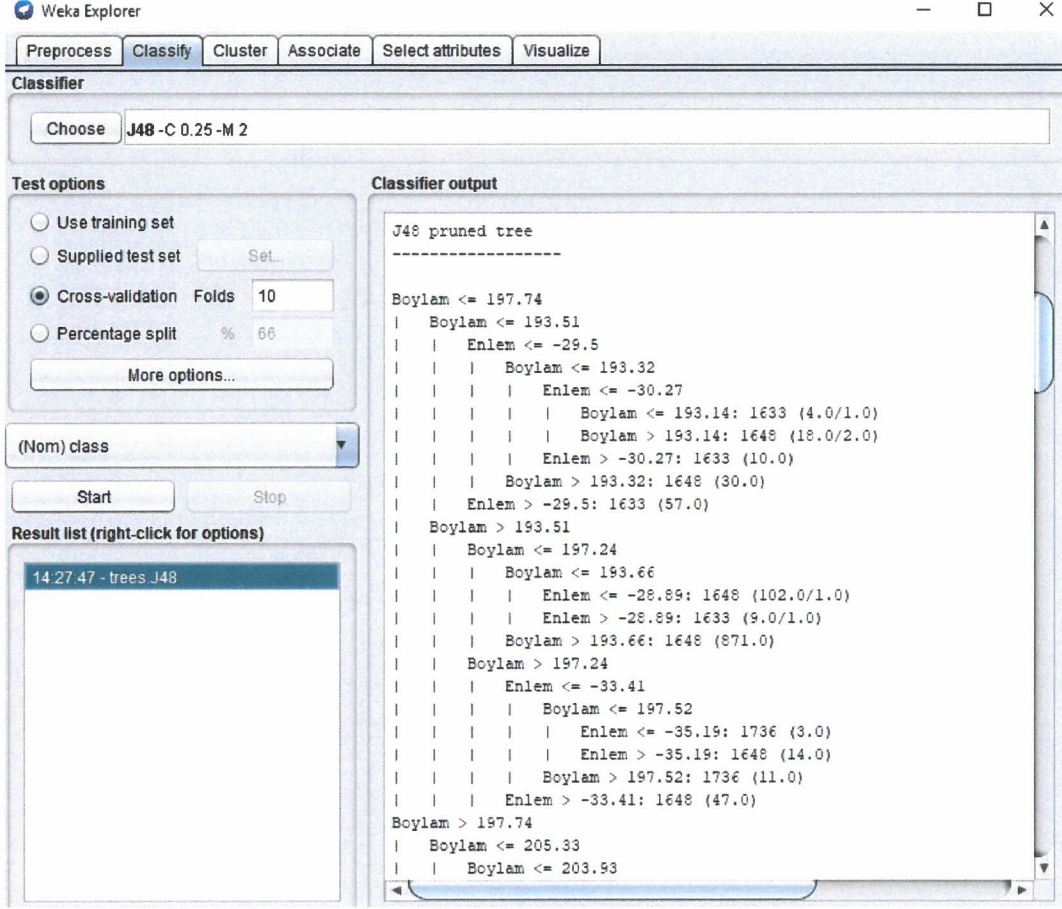
Oluşturulan modelin tahmin sonuçları ile gözlemlerin gerçek sınıflarının karşılaştırıldığı matris Çizelge 5.7’ de verilmiştir. Burada Abell 1633 sınıfı için model, 81 gözlemin 6 tanesini doğru tahmin edip 75 tanesini Abell 1648 grubunda tahmin etmiştir. Benzer şekilde Abell 1857 grubunda 26 gözlemden sadece 5 tanesini doğru tahmin etmiştir. Modelin diğer gruplardaki tahmin başarısı daha yüksektir.

Çizelge 5. 7. Çok Katmanlı Ağ Sınıflandırma Yöntemi için Karmaşıklık Matrisi

Gerçek Sınıf	Tahmin Sınıf							
	A1633	A1648	A1736	A1771	A1802	A1816	A1846	A1857
A1633	6	75	0	0	0	0	0	0
A1648	0	1086	8	0	0	0	0	0
A1736	0	3	2038	3	2	0	0	0
A1771	0	0	7	31	12	0	0	0
A1802	0	0	3	4	688	0	0	0
A1816	0	0	0	0	4	160	1	0
A1846	0	0	0	0	0	6	52	0
A1857	0	0	0	0	0	0	21	5

5.2.5 J48

J48 algoritması uygulanmasında dallanmadaki karmaşıklığı önlemek ve aşırı öğrenmeyi engellemek amacıyla budama kullanılmıştır. Weka ortamında budama işleminde budama güven faktörü 0.25 alınarak eğitim gerçekleştirilmiştir. Eğitim sonrasında elde edilen karar ağacı modelinin sadece bir parçası Şekil 5. 4’de verilmiştir. Budama yapılarak oluşturulan modelde bilgi kazanımı en yüksek olarak belirlenen boylam değişkeni kök düğüm olarak atanmıştır. İkinci dallanmada bilgi kazanımı en yüksek olan değişken yine boylam iken üçüncü dallanmada enlem olmuştur. Çizelge 5.8’de J48 sınıflandırma yönteminin tahmin sonuçları verilmiştir. Bu sonuçlara göre Abell 1857 sınıfına ait 26 gözlemin 19 tanesini doğru tahmin edilmiştir. 7 gözlemi ise Abell 1846 grubuna atanmıştır. Sonuçlar incelendiğinde modelin tahmin başarısının genel olarak yüksek olduğu görülmüştür. Modelin başarı yüzdesi %97.5 iken Kappa değeri 0.96 olarak bulunmuştur. Doğruluk, Duyarlılık ve Hassasiyet değerleri ise sırasıyla 0.99, 0.88 ve 0.90 olarak elde edilmiştir.



Şekil 5. 4. J48 yönteminden elde edilen modelin bir parçası

Çizelge 5. 8. J48 Sınıflandırma Yöntemi için Karmaşıklık Matrisi

Gerçek Sınıf	Tahmin Sınıf							
	A1633	A1648	A1736	A1771	A1802	A1816	A1846	A1857
A1633	71	10	0	0	0	0	0	0
A1648	8	1082	4	0	0	0	0	0
A1736	0	7	2032	3	4	0	0	0
A1771	0	0	6	37	7	0	0	0
A1802	0	0	5	4	670	16	0	0
A1816	0	0	0	0	10	150	5	0
A1846	0	0	0	0	0	5	50	3
A1857	0	0	0	0	0	0	7	19

5.2.6. LMT

Lojistik regresyon ve karar ağacı sınıflandırmasının birleşimi olan LMT sınıflandırma yönteminde düğümlerin bölüneceği minimum örnek sayısı 15 ve logit boost algoritmasının düzeltilmesi için ağırlık vektör değeri 0.01 olarak belirlendi. Oluşturulan modelin galaksilerin doğal gruplarını tahmin etme başarısı % 98.32, Kİ değeri 0.98 olarak bulunmuştur. Doğruluk 0.99, Duyarlılık 0.91 ve Hassasiyet değeri ise 0.97 olarak elde edilmiştir. Çizelge 5.9’da LMT sınıflandırma yönteminin tahmin sonuçlarının gerçek sınıflarla karşılaştırmalı sonuçları verilmiştir. Abell 1736 sınıfındaki 2046 gözlemin 2044 tanesi doğru, 2 tanesi farklı sınıflarda tahmin edilmiştir. Abell 1771 sınıfındaki 50 gözlemin sadece 28 tanesi doğru sınıflandırılmıştır. LMT sınıflandırma ile Abell 1633 sınıfındaki tüm gözlemler doğru tahmin edilmiştir.

Çizelge 5. 9. Sınıflandırma Yöntemi için Karmaşıklık Matrisi

Gerçek sınıf	Tahmin Sınıfı							
	A1633	A1648	A1736	A1771	A1802	A1816	A1846	A1857
A1633	81	0	0	0	0	0	0	0
A1648	2	1089	3	0	0	0	0	0
A1736	0	1	2044	1	0	0	0	0
A1771	0	0	13	28	9	0	0	0
A1802	0	0	7	0	668	0	0	0
A1816	0	0	0	0	3	160	2	0
A1846	0	0	0	0	0	1	55	2
A1857	0	0	0	0	0	0	3	23

5.2.7. Rasgele Orman

Birden fazla karar ağacından meydana gelen RO sınıflandırma yönteminde model oluşturmak için kullanılacak ağaç sayısı 100 olarak belirlendi. Modelin sınıflandırma başarısı % 98.14 olurken Kİ değeri de 0.97 olarak bulundu. Doğruluk, Duyarlılık ve Hassasiyet değerleri de sırasıyla, 0.99, 0.89 ve 0.95 olarak elde edilmiştir. Çizelge 5.10’da RO yönteminin sınıflandırma sonuçları verilmiştir. Oluşturulan model ile Abell 1736 sınıfının 2039 gözlemi doğru tahmin edilirken 5 gözleminin sınıfları farklı tahmin edilmiştir.

Çizelge 5.10. Rasgele Orman Sınıflandırma Yöntemi için Karmaşıklık Matrisi

Gerçek Sınıf	Tahmin Sınıfı							
	A1633	A1648	A1736	A1771	A1802	A1816	A1846	A1857
A1633	65	16	0	0	0	0	0	0
A1648	5	1085	4	0	0	0	0	0
A1736	0	4	2039	1	2	0	0	0
A1771	0	0	9	34	7	0	0	0
A1802	0	0	4	1	682	8	0	0
A1816	0	0	0	0	5	158	2	0
A1846	0	0	0	0	0	5	52	1
A1857	0	0	0	0	0	0	4	22

6. SONUÇLAR ve TARTIŞMA

Bu çalışmada Shapley Konsantrasyon bölgesine ait galaksilerin doğal sınıflarını tahmin etmek amacıyla öğrenmeye dayalı sınıflandırma modelleri geliştirilmiştir. Danışmanlı öğrenmeye dayalı Bayes Ağ, Naive Bayes, ÇKA, DVM, J48, LMT ve RO sınıflandırma yöntemleri Weka da Explorer ara yüzü kullanılarak sınıflandırılmıştır. Bu tez bu yönüyle literatürdeki çalışmalardan daha kapsamlı bir çalışma olmuştur.

Sınıflandırma yöntemlerinin performans kriterlerine göre karşılaştırması Çizelge 6. 1.'de verilmiştir. Yöntemlerin sınıflandırma başarı yüzdeleri genel olarak yüksektir. Karar ağaçları sınıflandırma yöntemlerinden LMT algoritmasının %98.88 oranıyla en iyi sınıflandırma yüzdesine sahip olduğu görülmüştür. Model tahmini ile gözlemin gerçek sınıfının uyumunu gösteren Kİ değerine bakıldığında LMT sınıflandırma yönteminin en yüksek uyuma sahip olduğu görülmektedir. Sonuçlara bakıldığında DVM sınıflandırma ile elde edilen modelin gerçek sınıfları tahmin etme başarısının düşük olduğu görülmektedir.

Çizelge 6. 1. Sınıflandırma Yöntemlerinin Performans Karşılaştırması

Yöntemler	BY	K	Doğruluk	Hassasiyet	Duyarlılık
Abell	100	1	1	1	1
Bayes Ağ	95.2788	0.9289	0.9882	0.7674	0.8419
Naive Bayes	94.2859	0.9217	0.9832	0.7765	0.8411
ÇKA	96.4650	0.9466	0.9912	0.7164	0.9237
DVM	94.0925	0.9101	0.9902	0.6019	0.6970
J48	97.5326	0.963	0.9938	0.8830	0.9041
LMT	98.8849	0.9832	0.9971	0.9712	0.9181
Rasgele Orman	98.1495	0.9722	0.9953	0.8940	0.9504

Doğruluk kriterine göre LMT sınıflandırmanın başarısı en yüksektir. Benzer şekilde Naive Bayes ve DVM sınıflandırma yöntemlerinin en düşük başarıya sahip olduğu görülmüştür. Duyarlılık ve hassasiyet kriterlerine göre de karar ağaçları sınıflandırma yöntemlerinin daha iyi sonuçlar verdiği sonucuna varılmıştır.

Yöntemlerin performansları arasındaki farklılıkların anlamlılığını test etmek için Friedman testi kullanılmıştır. Friedman testinin hipotezleri aşağıda verilmiştir.

H₀: Yöntemlerin performansları açısından istatistiksel olarak anlamlı farklılık yoktur.

H_A: Yöntemlerin performansları açısından istatistiksel olarak anlamlı farklılık vardır.

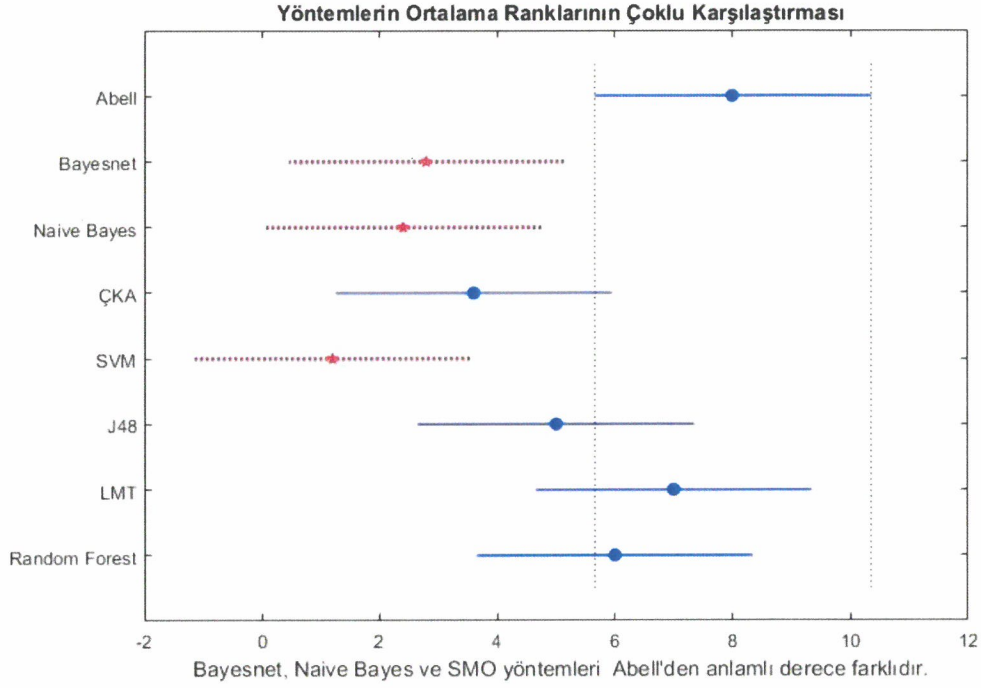
Friedman testi Matlab programlama da yapılmıştır. Anlamlılık düzeyi $p < 0.05$ olarak alınmış ve test sonuçları Çizelge 6. 2’de verilmiştir.

Çizelge 6. 2. Friedman Test Sonuçları

	SS	df	MS	Ki-kare	Olasılık(p) > Ki-kare
Sütun	198.4	7	28.3429	33.07	2.573e-05
Hata	11.6	28	0.4143		
Toplam	210	39			

Çizelge 6. 2’de SS, karelerin toplamı, df serbestlik derecesini, MS ise ortalama kare değerlerini göstermektedir. Ki-kare test istatistiği ve p değeri anlamlılık seviyesine göre testin gücü $p < 0.05$ olduğundan dolayı H_0 hipotezi reddedilir. Bu durum Shapley bölgesindeki galaksileri sınıflandırmak için kullanılan yöntemlerin performansları arasında anlamlı bir farklılık olduğunu ifade etmektedir. Ancak hangi yöntemlerin farklı olduğunun cevabını vermemektedir. Bunun için çoklu karşılaştırma testleri kullanılmıştır.

Sıra ortalamalarını karşılaştıran Friedman çoklu karşılaştırma testlerinin sonuçları Şekil 6.1’deki grafikte görülmektedir. Yöntemlerin performansları Abell baz alınarak değerlendirilmiştir. Grafikteki dikey kesikli çizgiler güven sınırlarını ifade etmektedir. Güven sınırları kesişen bölgeler birbirinden anlamlı derece farklı değildir. Yani Abell’in güven sınırları içerisinde kalan yöntemlerin Abell ile arasında anlamlı bir farklılık yoktur. KA yöntemleri ve ÇKA’nın Abell gruplarını tahmin etme de başarılı yöntemler olduğu görülmüştür. Diğer yöntemler ise Abell’den anlamlı derecede farklı olduğundan galaksi sınıflandırmasında başarılı tahmin yapamadıkları tespit edilmiştir.



Şekil 6. 1.Yöntemler arasındaki anlamlı farklılığın grafiksel gösterimi

Yöntemlerin performansları ile Abell gruplarının konumları ve birbirine uzaklıkları incelendiğinde şu sonuçlar elde edilmiştir. Abell gruplarının birbirine olan Öklid uzaklıkları Çizelge 5.1’de verilmiştir. Bu sonuçlara göre Abell 1633 ile Abell 1648 grubu ve Abell 1857 ile Abell 1846 gruplarının birine en yakın gruplar olduğu görülmektedir. Genel olarak bazı sınıflandırma modellerinin birbirine yakın sınıfları ayırmada başarılı olamadığı gözlenmiştir.Özel olarak da şu sonuçlar elde edilmiştir:

- *LMT sınıflandırma yöntemi* ile oluşturulan model en iyi sınıflandırma başarısına sahiptir. Diğer yöntemler 1633 grubundaki tüm gözlemleri doğru sınıflandıramamış olmasına rağmen LMT bu gruptaki tüm gözlemleri doğru sınıflandırmayı başarmıştır. Hatta DVM bu gruptaki gözlemlerin hiçbirini doğru tahmin edememiştir. Benzer şekilde LMT Abell 1857 grubundaki gözlemleri diğer yöntemlere göre en küçük hata ile tahmin etmiştir.
- LMT sınıflandırmadan sonra en iyi performansa sahip olan yöntem RO’dur. Bu yöntemle oluşturulan model birbirine yakın grupları ayırmada başarılı bir

performansa sahiptir.

- *J48 sınıflandırma yöntemi* RO sınıflandırmadan sonraki en yüksek başarıya sahip olan yöntemdir. Ayrıca bu yöntem ile sınıfları ayırmada en etkili değişkenin boylam değişkeni olduğu tespit edilmiştir.
- KA yönteminden sonraki en başarılı sınıflandırma modeli ÇKA ile oluşturulan modeldir. Bu modelin birbirinden uzak grupları ayırma da daha başarılı olduğu görülmüştür. Birbirine yakın olan gruplarda hata oranı yüksektir.
- *Bayes Ağları*, birbirine yakın olan grupları doğru ayıramazken birbirine uzak grupları da düşük başarı ile tahmin etmiştir. Ancak Naive Bayes ve DVM yöntemlerinden daha başarılıdır.
- *Naive Bayes sınıflandırma* ile oluşturulan model birbirine yakın veya uzak grupları ayırt etmekte başarılı bir performans göstermemiştir.
- *DVM yöntemi ile* oluşturulan modelin tüm yöntemler arasında sınıflandırma hatası en yüksek olan modeldir. Ayrıca Abell 1633 grubundaki hiçbir gözlemi doğru tahmin edememiştir.

Sonuç olarak Bayes Ağ, Naive Bayes, DVM yöntemleri birbirine yakın olan grupları ayırmada başarılı olamamıştır.

Birbirine yakın sınıfları tahmin etmede en başarılı yöntem KA içerisinde LMT'dir.

KA ve ÇKA modellerinin galaksilerin doğal gruplarını tespit etme başarısı diğer yöntemlerden daha yüksektir. Ancak yöntemler kendi aralarında karşılaştırıldığında performans sıralaması LMT, RO, J48 ve ÇKA şeklinde gerçekleşmiştir. Literatürde genellikle Naive Bayes, DVM ve RO yöntemleri karşılaştırılmıştır ve en iyi sonucu RO vermiştir. Bu çalışma da farklı olarak 7 ayrı sınıflandırma yöntemi kullanılmıştır ve bu yöntemler arasında en yüksek performansı yine KA yöntemlerinden LMT göstermiştir. Bu sonuçlar galaksi sınıflandırmasında diğer yöntemlere alternatif olarak LMT sınıflandırmanın da kullanılabileceğini göstermiştir.

Sonuç olarak galaksi sınıflandırılmasında KA yöntemleri ve ÇKA kullanılması önerilebilir. Bu yöntemler ile oluşturulan modellerin Shapley Konsantrasyon bölgesindeki galaksilerin doğal gruplarını tahmin başarısının yüksek olması, bu çalışmanın farklı bölgeler için de genişletilebileceğini göstermesi açısından önemlidir. Teknolojik

geliřmelerle birlikte evrenin derinliklerinde neredeyse her gn yeni galaksiler veya gk cisimleri keřfedilmektedir.

Bu yeni galaksilerin sınıflandırılmasında, daha nce bilinen gk cisimlerine ilişkin verilerinin ğrenme yntemleriyle blgesel sınıflandırma modellerinin yaratılması arařtırmacılara zaman ve maliyet aısından byk avantajlar saėlayabilir. İleri ki alıřmalarda bu koordinatlar geniřletilerek tm gkyz katalogundaki (SDSS) galaksi sistemlerine uygulanması amalanmaktadır.

KAYNAKLAR

- Abell, G. O., (1957) The Distribution of rich clusters of galaxies, *The Astrophysical Journal Supplement Series*, 3: 211-288.
- Abell, G. O., Corwin, H. G., Jr., Olowin, R. P., (1988), A Catalog of Rich Clusters of Galaxies, *The Astrophysical Journal Supplement Series*, 70: 1-138.
- Aydın, E. (2011), *Yapay Öğrenme*, Boğaziçi Üniversitesi Yayinevi, 486s.
- Bailin J. ve Harris W.E., (2008), "Inclination-Independent Galaxy Classification", *The Astrophysical Journal*, 681: 225-231
- Breiman L. 2001, "Random Forests", *Machine Learning*, 45(1): 5-32.
- Canküyer E., Aşan E. (2005) *Parametrik Olmayan İstatistiksel Teknikler*, Anadolu Üniversitesi Fen Fakültesi Yayınları, Eskişehir, 259s.
- Cheng, T. Y., Conselice, C. J., Aragón-Salamanca, A., Li, N., Bluck, A. F., Hartley, W. G., & James, D. J. (2019). Optimising Automatic Morphological Classification of Galaxies with Machine Learning and Deep Learning using Dark Energy Survey Imaging, *Instrumentation and Methods for Astrophysics (astro-ph.IM)*, 1908.03610
- Cortes C., Vapnik V., (1995), Support-Vector Networks, *Kluwer Academic Publishers Machine Learning*, 20 (3), 273-297
- J. C. Gower, (1982), Euclidean distance geometry, *Math. Sci.* 1: 1-14
- Dobrycheva, D. V., Vavilova, I. B., Melnyk, O. V., & Elyiv, A. A. (2017). Machine learning technique for morphological classification of galaxies at $z < 0.1$ from the SDSS. *arXiv preprint arXiv:1712.08955*.

- Domingos, P., and Pazzani, M., (1997) Beyond independence: Conditions for the optimality of the simple Bayesian classifier, *Machine Learning* 13: 105-115.
- Dressler (1979), Catalog Of Morphological Types In 55 Rich Clusters Of Galaxies, *The Astrophysical Journal Supplement Series*, 42: 565-609
- Driver S.P., Liske J., Cross N. J. G., De Propriis R. and Allen P. D., (2005), The Millennium Galaxy Catalogue: The Space Density And Surface-Brightness Distribution(S) Of Galaxies, *Mon. Not. R. Astron. Soc.* 360: 81–103.
- Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the american statistical association*, 32(200): 675-701.
- Gauci A., Kristian Zarb Adami K.Z. and Abela J., (2010), Machine Learning for Galaxy Morphology Classification, *Mon. Not. R. Astron. Soc.*, 000 1–8.
- Gauthier A., Archa Jain A. and Emil Noordeh E., (2016), Galaxy Morphology Classification, *Stanford University*. (Ders Notu)
- Godbole S., Sarawagi S., and Chakrabarti S., (2002) Scaling multi-class support vector machines using inter-class confusion, *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. 513-518
- Goderya N.S. and Lolling S.M., (2001), Morphological classification of Galaxies Using Computer Vision and Artificial Neural Networks: A Computational scheme, *Astrophysics and Space Science* 279: 377–387.
- Han J., Kamber M., Pei J., (2006), *Data Mining : Concepts and Techniques*, Morgan Kaufmann Publishers, Third Edition, Elseiver, 740s.
- Haykin S., 2009, *Neural Networks and Learning Machines*, Pearson. Upper Saddle River, NJ, USA, 938s.
- Hubble E., (1926) Extra-galactic Nebulae, *Contributions from the Mount Wilson Observatory / Carnegie Institution of Washington*, 324: 321-369

- J. C. Gower, (1982), Euclidean distance geometry, *Math. Sci.* 1:1-14
- N. Landwehr, M. Hall, E. Frank, (2005), Logistic Model Trees, *Machine Learning*, 59: 161-205
- Lewis, D. D., (1998) Naive (Bayes) At Forty: The Independence Assumption In Information Retrieval, *Springer Berlin Heidelberg*, 98: 4-15 1998
- Lotz, J.M., Primack, J., Madau, P., (2004), A New Nonparametric Approach To Galaxy Morphological Classification, *The Astronomical Journal*, 128(1): 163–182.
- Marin.M. v.d., 2013, “A Hierarchical Model for Morphological Galaxy Classification”, Proceedings of the Twenty-Sixth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2013, St. Pete Beach, Florida, USA, May 22-24.
- Miller A. S. and Coe M. J. (1996), Star/Galaxy Classification Using Kohonen Self-Organizing Maps, *Mon. Not. R. Astron. Soc.* 279: 293-300
- Tan, P.N., Steinbach M., Kumar V., (2006), *Introduction to Data Mining*, Pearson Education. Indiana: Addison-Wesley, 169s.
- Quinlan T.J.R, (1986), Induction of Decision, *Kluwer Academic Publishers Machine Learning* 1: 81-106.
- Paul, N., Virag, N., & Shamir, L. (2018), A catalog of photometric redshift and the distribution of broad galaxy morphologies. *Galaxies MDPI*, 6(2): 64.
- Remya G. and Mohan A., (2016), Deep Learning Approach for Classifying Galaxies, 6(4): 823-828
- Sandage, A., (1961), *Hubble Atlas of Galaxies*, Carnegie Institution of Washington, Washington D.C, 50s.
- Schilling G. (2014) *Derin Uzay Tübitak Popüler Bilim Kitapları*, Salmat Basım

Yayıncılık, Ankara, 224s.

Schneider, P., (2006), *Extragalactic Astronomy and Cosmology an Introduction*, Springer Yayinevi, Berlin. 474s.

Selim, I., Keshk, A. E., & El Shourbugy, B. M. Galaxy image classification using non-negative matrix factorization. *International Journal of Computer Applications*, 137(5), 4-8, 2016.

Shapley, H. (1930), Note on a Remote Cloud of Galaxies in Centaurus, *Harvard College Observatory Bulletin*, 874: 9-12

Sparke, L.S., Gallagher, J.S., (2007), *Galaxies in the Universe: An Introduction*, Cambridge University Press, New York, 421s.

Stigler, S. M. (1982), Thomas Bayes Bayesian Inference, *Journal of the Royal Statistical Society, series* 145: 250-258

Vaucouleurs, G. D., (1959), Classification and Morphology of External Galaxies, *Handbuch der Physik*, 53(11): 275-310

Wampold, B. E., (1989), Kappa As A Measure Of Pattern In Sequential Data, *Quality&Quantity*, 23: 171-187.

White, H. (1989). Learning in artificial neural networks: A statistical perspective. *Neural computation*, 1(4), 425-464.

http 1: https://astrostatistics.psu.edu/datasets/Shapley_galaxy.dat

http 2: <http://vizier.u-strasbg.fr/viz-bin/VizieR-2>

http 3: <https://www.eso.org/public/images/eso9845d/>

http 4: <https://www.eso.org/public/images/eso1516b/>

http 5: <https://www.eso.org/public/images/ngc/>

ÖZGEÇMİŞ

Kişisel Bilgiler

Ad Soyad :Nazlı Deniz ERGÜÇ
Uyruk : T.C.
Doğum Yeri ve Tarihi: Hatay- 23/09/1993
Medeni Hali : Bekar
Telefon : 0 544 440 9788
E-posta : ndenizerguc@gmail.com

Eğitim

Alınan Derece	Aldığı Kurum/Üniversite	Mezuniyet Yılı
Lise	Reyhanlı Lisesi	2010
Lisans	Çukurova Üniversitesi	2015
Yüksek Lisans	Muğla Sıtkı Koçman Üniversitesi	2019

İş Tecrübesi

Yıl	Yer	Pozisyon/görev
2019-	Çağlar Teknoloji SAN. TİC. A.Ş.	Denetmen

Yabancı Dil(ler)

Dil (İngilizce, vs)	Başlangıç	Orta	İleri
Yazma	X		
Konuşma	X		
Anlama		X	
Okuma		X	

Ödüller

Muğla Sıtkı Koçman Üniversitesi “8. Fen Bilimleri Araştırma Sempozyumu” Poster Yarışması: Sıtkı Koçman *Birincilik* Poster Ödülü

Bilimsel Faaliyetler

Nida GÖKÇE NARİN, Nazlı Deniz ERGÜÇ "Comparison Of Classification Performance Of The Galaxies in The Shapley Concentration Region", 11. ISDC 2018

Nazlı Deniz ERGÜÇ, Nida GÖKÇE NARİN, “Shapley Konsanstrasyon Bölgesindeki Galaksilerin İstatistiksel Öğrenme Yöntemleriyle Sınıflandırılması” 8. Fen bilimleri Araştırma Sempozyumu, 2018

Yayınlar

Nazlı Deniz ERGÜÇ, Nida GÖKÇE NARİN, Classification Of Galaxies In Shapley Concentration Region With Machine Learning”, Mugla Journal of Science and Technology, (2019)