DOKUZ EYLÜL ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

TOPLULUK ÖĞRENMESİ YAKLAŞIMIYLA ANOMALİ BELİRLEME

Nihat AKILLI

Mart, 2023
İZMİR

TOPLULUK ÖĞRENMESİ YAKLAŞIMIYLA ANOMALİ BELİRLEME

Dokuz Eylül Üniversitesi Fen Bilimleri Enstitüsü
Yüksek Lisans Tezi
İstatistik Anabilim Dalı, Veri Bilimi Programı

Nihat AKILLI

Mart, 2023
İZMİR

YÜKSEK LİSANS TEZİ SINAV SONUÇ FORMU

Nihat AKILLI tarafından, **Dr. Öğr. Üye. Engin YILDIZTEPE** yönetiminde hazırlanan **"TOPLULUK ÖĞRENMESİ YAKLAŞIMIYLA ANOMALİ BELİRLEME"** başlıklı tez tarafımızdan okunmuş, kapsamı ve niteliği açısından Yüksel Lisans tezi olarak kabul edilmiştir.

Dr.Öğr.Üyesi Engin YILDIZTEPE

Yönetici

Doç.Dr. Hakan Savaş SAZAK

Jüri Üyesi

Dr.Öğr.Üyesi A.Övgü KINAY

Jüri Üyesi

Prof. Dr. Okan FISTIKOĞLU
Müdür
Fen Bilimleri Enstitüsü

TEŞEKKÜR

Yol göstericiliği ve katkılarıyla akademik alanda gelişimime destek sağlayan ve beni motive eden Veri Bilimi Yüksek Lisans Programı tez danışmanım Dr. Engin YILDIZTEPE'ye değerli katkılarından dolayı; her zaman yanımda olan çok kıymetli eşim Gizem'e sabır ve anlayışından ötürü çok teşekkür ederim.

Nihat AKILLI

TOPLULUK ÖĞRENMESİ YAKLAŞIMIYLA ANOMALİ BELİRLEME

ÖZ

Veri analizlerinde anomaliler bazen veriden çıkartılması gereken kayıt hataları olabilecekleri gibi, bazen de tespit edilmesi hayati öneme sahip değerler olabilir. Bu sebeple anomali belirlemeye farklı uygulama alanlarında ihtiyaç duyulmaktadır. Çoğunluğun kararının daha isabetli olacağı prensibiyle uygulanan topluluk öğrenmesi, sınıflama çalışmalarında tercih edilen bir yaklaşımdır. Topluluk öğrenmesi yaklaşımıyla anomali belirleme ise veri bilimi alanında ilgi gören güncel bir çalışma alanıdır. Anomalilerin nadir gerçekleştiği varsayımı altında, denetimli anomali belirleme, topluluk öğrenmesi yaklaşımıyla çalışan sınıflayıcılar için bir dengesiz sınıflandırma problemi olarak ele alınabilir. Bu tezde, topluluk öğrenmesi yaklaşımlarının uygulandığı sınıflama yöntemlerinin anomali belirlemede kullanılması araştırılmıştır. Çalışmada, yöntemlerin sınıf dengesizliğini gideren teknikler ile birlikte kullanılması durumu da incelenmiştir. Yapılan uygulamalar sonucunda, sınıf dengesizliğini gideren yöntemler ile birlikte kullanıldığında topluluk öğrenmesi yaklaşımlarının anomali belirleme problemlerinde başarılı sonuçlar verdiği görülmüştür.

Anahtar Kelimeler: Anomali belirleme, topluluk öğrenmesi, sınıf dengesizliği, ADASYN, ROS, Boosting, Bagging

ANOMALY DETECTION WITH ENSEMBLE LEARNING APPROACH

ABSTRACT

In data analysis, anomalies can sometimes be recording errors that need to be removed from the data, and at other times they can be crucial values to detect. For this reason, anomaly detection is essential in various application areas. Ensemble learning, which uses the principle that the majority's decision will be more accurate, is a preferred approach in classification studies. The use of ensemble-based techniques for anomaly detection is a current area of study in data science. Given the assumption that anomalies are rare, supervised anomaly detection can be treated as an imbalanced classification problem for classifiers utilizing ensemble learning. This thesis investigates the use of classification methods that incorporate ensemble learning approaches for anomaly detection. The study also examines the use of techniques that eliminate the class imbalance problem in conjunction with these methods. As a result of the applications, it has been found that ensemble learning approaches are successful in anomaly detection problems when used in conjunction with methods that eliminate class imbalance.

Keywords: Anomaly detection, ensemble learning, class imbalance, ADASYN, ROS, Boosting, Bagging

İÇİNDEKİLER

Sayfa

ŞEKİLLER LİSTESİ

Sayfa

TABLOLAR LİSTESİ

Sayfa

KISALTMALAR

RUS	: Random Under Sampling (Rastgele Alt Örnekleme)
ROS	: Random Over Sampling (Rastgele Aşırı Örnekleme)
SMOTE	: Synthetic Minority Over-Sampling Technique
ADASYN	: Adaptive Synthetic Sampling
CT	: Classification Tree (Sınıflandırma Ağacı)
CART	: Classification and Regression Tree
RF	: Random Forest
GBM	: Gradient Boosting Machines
XGBoost	: Extreme Gradient Boosting
AdaBoost	: Adaptive Boosting
iForest	: Isolation Forest
ROC	: Receiver Operating Characteristics
AUC	: Area Under Curve
PR	: Precision-Recall
DP	: Doğru Pozitif
DN	: Doğru Negatif
YP	: Yanlış Pozitif
YN	: Yanlış Negatif
TM	: Toplam Maliyet
XG	: XGBoost

BÖLÜM BİR
GİRİŞ

Bilgi teknolojilerinin gelişmesiyle dünyada üretilen veri katlanarak artmaktadır. Bu verilerden anlamlı bilgi çıkarmayı amaçlayan yapay zekânın önemli uygulama alanlarından biri de makine öğrenmesidir. Birden fazla makine öğrenmesi yönteminin birlikte kullanılmasıyla gerçekleştirilen öğrenmeye ise topluluk öğrenmesi yaklaşımı denilmektedir. Çoğunluk veri içerisinde geneli yansıtmayan alışılmışın dışındaki gözlemlerin tespit edilmesi süreci ise anomali belirleme olarak tanımlanabilir. Topluluk öğrenmesi yaklaşımları ile anomali belirleme son yıllarda makine öğrenmesi alanında araştırma yapılan alanlardan birisi olmuştur.

Veri içerisindeki çoğunluktan ayrılan tek bir gözlem, bir örüntü veya bağlamsal anomali olarak değerlendirilen durumların belirlenmesi, finans alanında sahte işlemlerin tespiti, bilgi sistemlerinde sistem hatalarının tespiti, bilgisayar ağlarında ağ saldırılarının tespiti, sağlık alanında hastalıklara teşhis konulması, endüstriyel sistemlerde hasar veya arıza tespiti gibi oldukça yaygın kullanım alanına sahiptir.

Anomali belirleme genellikle üç farklı yaklaşım ile tanımlanır. Verideki gözlemlerin olağan ve anomali olarak nitelendirildiği etiketli veriler kullanılarak eğitilen modeller ile anomali belirleme denetimli anomali belirleme olarak isimlendirilir. Sadece olağan gözlemlerin nitelenerek bu gözlemlerden farklı az sayıda gözlemle karşılaşıldığında bunların anomali olarak ayırt edilmesine ise yarı denetimli anomali belirleme denilmektedir. Bir etiket ile olağan veya anomali olarak nitelendirilmemiş verileri kullanan, benzer yapıdaki gözlemlerin değerlendirilmesi ile anomalilerin belirlendiği yaklaşıma ise denetimsiz anomali belirleme denir.

Hayatta karşılaşılan sorunlarla ilgili bir tercih yapmak veya önemli bir karar vermek gerektiğinde birden çok kişiden fikir alarak çoğunluğun kararının bir kişinin görüşünden daha isabetli olacağı değerlendirilir. Bu değerlendirmenin makine öğrenmesindeki karşılığı topluluk öğrenmesidir. Topluluk öğrenmesi makine öğrenmesinin son yıllarda oldukça popülerlik kazanmış bir uygulamasıdır. Topluluk

öğrenmesi ile, birden fazla yöntem bir arada kullanılarak, bir algoritma farklı parametreler ile çalıştırılarak veya veri kümelerinin alt kümeleri kullanılarak elde edilen sonuçların birlikte değerlendirilmesiyle daha isabetli tahminleme yapılması, daha güçlü bir genelleme kabiliyeti amaçlanmaktadır.

Topluluk öğrenmesi yaklaşımı, anomali belirleme problemlerinde üç başlık altında incelenmektedir. Etiketli veride sınıflar arasında oransal olarak oldukça yüksek fark bulunduğu durumlarda denetimli topluluk öğrenmesi yöntemleriyle sınıflama işlemi anomali belirleme olarak değerlendirilebilir. Bu yaklaşım, topluluk öğrenmesi ile denetimli anomali belirleme olarak adlandırılır. Sadece normal gözlemlerden oluşan bir veri ile oluşturulan modelin farklı nitelikte az sayıda gözlem ile karşılaştığında bunu anomali olarak ayırt ettiği durum ise topluluk öğrenmesi yaklaşımı ile yarı denetimli anomali belirleme olarak adlandırılır. İçerdiği gözlemlerden büyük çoğunluğunun benzer özellikler taşıdığı varsayılan veri etiketi içermeyen bir veri kümesinde topluluk öğrenmesi yöntemleriyle analiz yapılarak küçük grupta yer alan gözlemlerin anomali olarak nitelendirilmesi yaklaşımı ise topluluk öğrenmesi ile denetimsiz anomali belirleme olarak adlandırılır.

Gerçek hayat uygulamalarında karşılaşılan problemlerde analize ihtiyaç duyulan verilerin içerdiği sınıflara göre dengesiz dağılması oldukça olası bir durumdur. Geleneksel sınıflayıcılar, dengesizlik durumunda çoğunluk sınıfı sayesinde yüksek doğruluk değerine ulaşma eğiliminde olmakta ancak azınlık sınıfını belirlemede beklenen performansı gösterememektedir. Sınıf dengesizliği problemini gidererek, topluluk öğrenmesi yaklaşımı ile denetimli anomali belirleme yöntemlerini uygulamak başarılı sonuçlar verebilmektedir (Yıldıztepe ve Akıllı, 2021). Sınıf dengesizliğini gidermede yeniden örneklemeye dayalı tekniklerin uygulanması belirtilen soruna getirilen çözümlerden biridir.

Literatürde topluluk öğrenmesi yöntemleri, anomali belirleme yöntemleri ve yeniden örnekleme yöntemleri üzerine çok sayıda çalışma bulunmaktadır. Ancak bu yöntemlerin birlikte kullanılması fikri yaygın değildir. Literatürde topluluk öğrenmesi yöntemleriyle birlikte yeniden örneklemeye dayalı çözümü araştıran çeşitli yöntemler

önerilmiştir. Chawla, Lazarevic, Hall ve Bowyer (2003) SMOTEBoost algoritmasını önermişlerdir. Bu yöntem ile en bilinen algoritmalardan olan SMOTE yöntemi ile en çok kullanılan topluluk öğrenmesi yöntemlerinden Boosting yöntemini birleştirmişler, Boosting yönteminde her aşamada veri kümesindeki gözlemlerden yanlış sınıflandırılan gözlemlerin ağırlıklarını değiştirmek yerine SMOTE algoritmasıyla yeni azınlık örnekler ekleyerek tahminleme yapmışlardır.

Seiffert, Khoshgoftaar, Hulse ve Napolitano (2008) RUSBoost algoritmasını önermişlerdir. Bu yöntem ile Boosting yöntemi ile Rastgele az örnekleme (RUS) yönteminin birleştirilmesi amaçlanmıştır. SMOTEBoost algoritmasıyla benzer mantığı olmakla birlikte eğitim süresi açısından daha başarılı olduğu belirtilmiştir.

Wang ve Yao (2009) tarafından önerilen OverBagging yönteminde Bagging yöntemi ile Rastgele aşırı örnekleme (ROS) yönteminin birleştirilmesi amaçlanmıştır. Yine benzer şekilde Hanifah, Wijayanto ve Kurnia (2015) tarafından SMOTE yöntemi ve Bagging yöntemini birlikte uygulandığı yönteme SMOTEBagging adını vermişlerdir.

Liu, Wu ve Zhou (2008) Easy Ensemble ve Balance Cascade isminde hem Bagging hem Boosting yönteminden yararlanılan birbirine benzer iki yeni yöntem önermişlerdir. Easy Ensemble yönteminde azınlık sınıfının tüm gözlemleri kullanılırken çoğunluk sınıftan yeniden örnekleme yapılarak alt örneklemler oluşturulur ve her bir alt örneklem için model oluşturularak çıktılar birleştirilir. Balance Cascade yöntemi ise ilerlemeli devam eder. Çoğunluk sınıfa ait gözlemlerden yüksek doğruluk ile sınıflandırılanlar sonraki aşamada doğru sınıflandıran sınıflandırıcılar tarafından değerlendirmeden çıkarılır ve sonraki aşamada bu gözlem yer almadan işlem tekrar ettirilir.

Topluluk öğrenmesi yaklaşımıyla anomali belirleme daha çok denetimli öğrenme yöntemleri ile gerçekleştirilen bir görevdir. Bu çalışmada denetimli öğrenme yöntemleri olarak Bagging, Boosting, Oylama ve Stacking yöntemlerine yer verilmiş, denetimsiz anomali belirleme yaklaşımı olarak Isolation Forest yöntemi ele alınmıştır.

Uygulama kısmında ise topluluk öğrenmesi ile denetimli anomali belirleme çalışılmıştır. Tez çalışması beş bölümden oluşmaktadır. İkinci bölümde anomalinin tanımına, anomali türlerine, anomali belirleme yaklaşımlarına yer verilmiştir. Üçüncü bölümde topluluk öğrenmesi yöntemleri ve bu yöntemlerle anomali belirleme incelenmiş, sınıf dengesizliği problemine çözüm yaklaşımlarına değinilmiş, sınıf dengesizliği problemine çözüm olarak önerilen yeniden örnekleme tabanlı yöntemlere, model performans değerlendirme ölçütlerine yer verilmiştir. Dördüncü bölümde topluluk öğrenmesi yaklaşımıyla denetimli anomali belirleme ile ilgili uygulama gerçekleştirilmiştir. Uygulama sonuçlarında yanılgı matrisi kullanılarak elde edilen AUC ve AUC-PR değerleri ile F_1 ve $F_{0.5}$ puanları dikkate alınarak Bagging ve Boosting topluluk öğrenmesi yöntemlerinin anomali belirlemede başarısı incelenmiştir. Tezin son bölümünde çalışmadan çıkarılan sonuçlar tartışılmış ve gelecek çalışmalara ilişkin yorumlara yer verilmiştir.

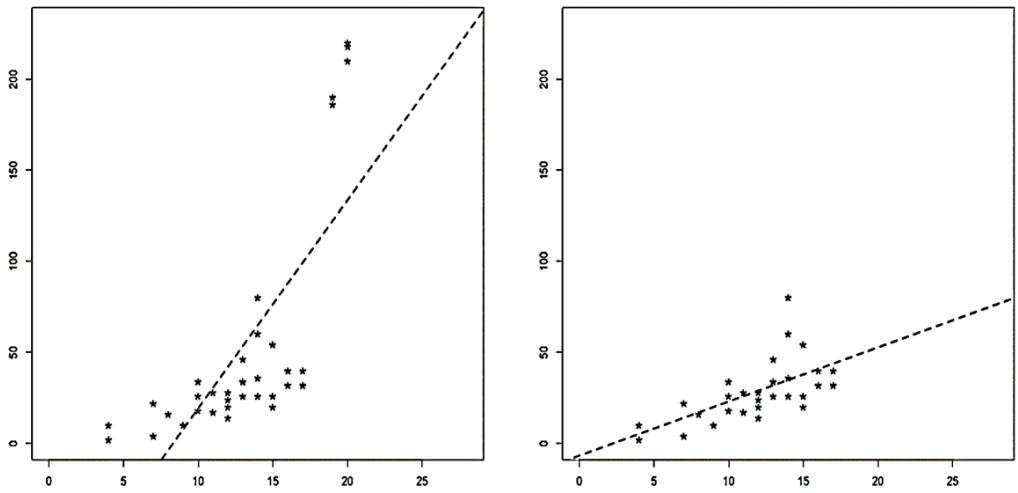
BÖLÜM İKİ
ANOMALİ BELİRLEME

Veri analizi, toplanan verinin amaca uygun şekilde kıymetlendirilerek işe yarayan bilginin çıkarılması sürecidir. Sağlıklı bir analiz için veri içerisindeki geneli yansıtmayan alışılmışın dışındaki gözlemlerin tespit edilerek ayrıştırılması veya bu gözlemlerin ihtiyaca göre değerlendirilmesi gerekir. Alışılmışın dışında olduğu değerlendirilen gözlemler duruma göre sağlıklı analizi engelleyen sorunlu gözlemler olarak görülürler ve analiz öncesi ayıklanarak dikkate alınmazlar. Bazen de esas incelenmesi dikkate alınması gereken gözlemler bu gözlemler olabilir. Literatürde bu kapsamdaki çalışmalar anomali belirleme veya aykırı değer tespiti başlıkları altında incelenmiştir. Bu çalışmada anomali belirleme başlığı tercih edilmiştir.

2.1 Anomali

Belirli bir alanda toplanan verilerin, özelliklerine göre ayrıştırılması sonucunda verinin içerisinde beklendiği gibi olmayan durum, gözlem ve desenlere anomali denir. Anomalilerin temel ayırt edici özellikleri veri içerisindeki çoğunluk gözlem ve desenlerden belirgin şekilde farklı olmalarıdır. Anomaliler, verinin çoğunluğunun uyduğu mekanizmadan farklı bir şekilde oluşmuş olabileceğini düşündürür (Chandola, Banerjee, ve Kumar, 2009).

Anomali, bazen yanlış bir ölçüm neticesinde ortadan kaldırılması gereken bir kayıt hatası olabileceği gibi, bazen de tespit edilmesi ve değerlendirilmesi oldukça hayati öneme sahip bir değer de olabilir (Wang, Bah, ve Hammad, 2019). Anomalinin tanımı çalışmanın içeriğine ve hedefine göre karar verilmesi gereken bir konudur.



Şekil 2.1 Anomali değerler çıkarılmadan ve çıkarılarak yapılan bir regresyon çalışması (Prabhakaran, 2017)

Şekil 2.1'de anomali değerler çıkarılmadan ve çıkarılarak yapılan gözlem sonuçlarına göre oluşturulan regresyon doğruları arasındaki fark gösterilmiştir. Bu örnekte anomali değerlerin regresyon modelinin doğru tahmin edilmesini etkilediği görülmektedir. Böyle durumlarda anomaliler bulunarak veriden temizlenmelidir.

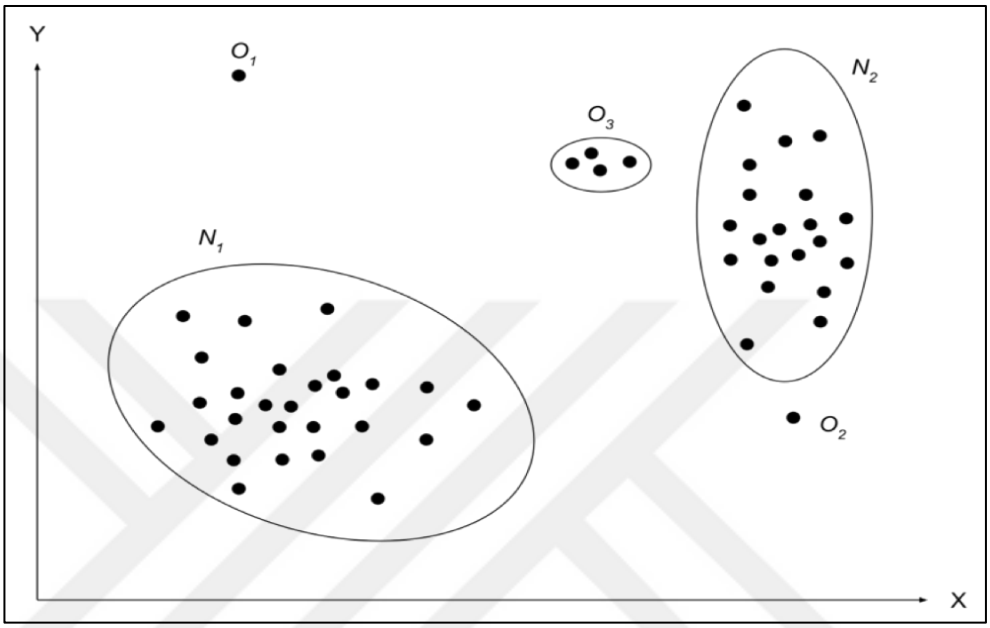
Ancak anomaliler her durumda sadece veriden çıkarılmak için tespit edilmez. Anomali belirleme, bankacılık alanında sahte işlem tespiti, kritik sistemlerde sistem hatalarının tespiti, bilgisayar ağlarında saldırı tespiti, sağlık alanında teşhis konulması, endüstride hasar veya arıza tespiti gibi oldukça yaygın kullanım alanına sahiptir (Wang vd., 2019).

2.2 Anomali Türleri

Anomaliler, nokta anomali, bağlamsal anomali, kolektif anomali olmak üzere üç alt kategoride incelenmektedir.

Verinin geri kalan çoğunluğuna göre farklılık gösteren tekil gözlemlere nokta anomali denir (Chandola vd., 2009). Şekil 2.2'deki örnekte O_1 noktasının diğer tüm gözlemlere göre belirgin şekilde ayrı durduğu görülebilmektedir. O_2 noktası ise tüm gözlemler incelendiğinde anomali olarak değerlendirilmeyebilir ancak N_2 kümesine

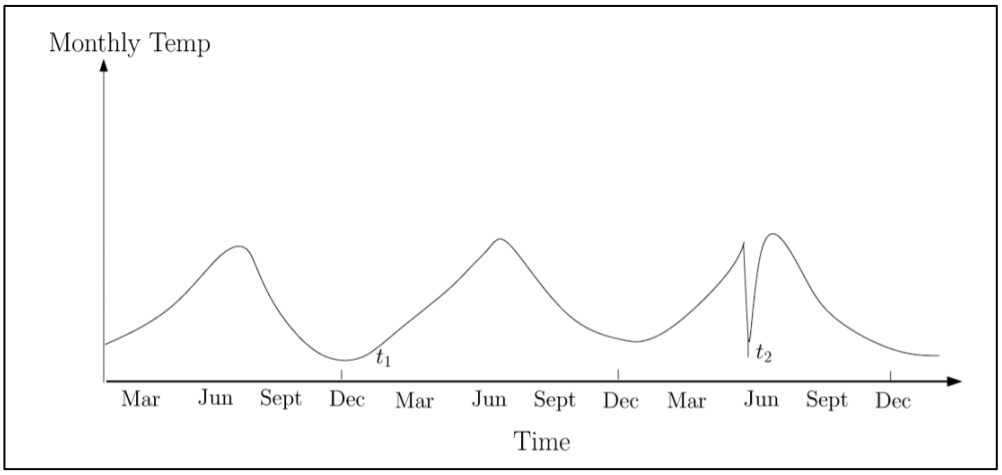
göre değerlendirildiğinde anomali olarak etiketlenebilir. Ayrıca O_3 veri kümesi de diğer tüm gözlemlere göre nokta anomali grubu olarak nitelendirilebilir. Bu örnekler, gözlemlerin anomali olup olmadığının subjektif bir değerlendirme olabildiğini ve alan bilgisine sahip araştırıcı tarafından belirlenmesi gerektiğini göstermektedir.



Şekil 2.2 Nokta anomali örneği (Chandola vd., 2009).

Tek bir gözlem olarak ele alındığında anomali olarak görülmeyen ancak verinin bulunduğu bağlama göre incelendiğinde anomali olarak değerlendirilen durumlara bağlamsal anomali denir (Chandola vd., 2009).

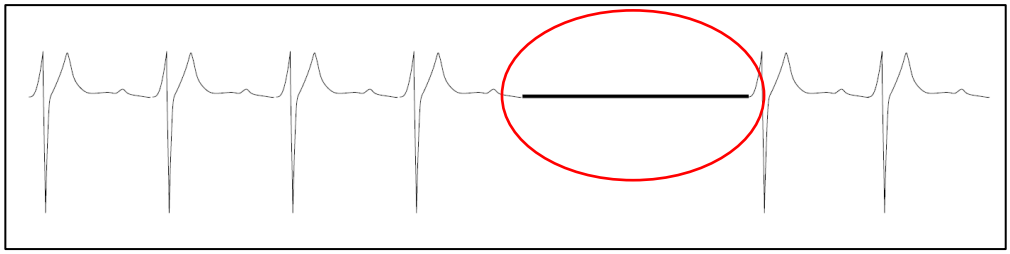
Örneğin, İzmir ilinde Aralık ayında 30 derece sıcaklık ölçülmesi bağlamsal bir anomalidir. Çünkü 30 derece İzmir için bahar ve yaz aylarında normal bir sıcaklık değeriyken kış aylarında bu sıcaklığın ortaya çıkması beklenmez. İzmir ilinde Aralık ayında daha düşük sıcaklıklar beklendiği için bu durum bağlamsal anomali olarak değerlendirilebilir. Bağlamsal anomalilere daha çok zaman serilerinde rastlanmaktadır.



Şekil 2.3 Bağlamsal anomali örneği (Chandola vd., 2009).

Şekil 2.3'te verilen örnekte, t_1 ve t_2 sıcaklık değerleri birbirine eşit olmasına rağmen bulundukları mevsim itibariyle sadece t_2 değerine sahip gözlemin kendisine zaman olarak yakın diğer gözlemlere göre bağlamsal anomali olduğu görülmektedir.

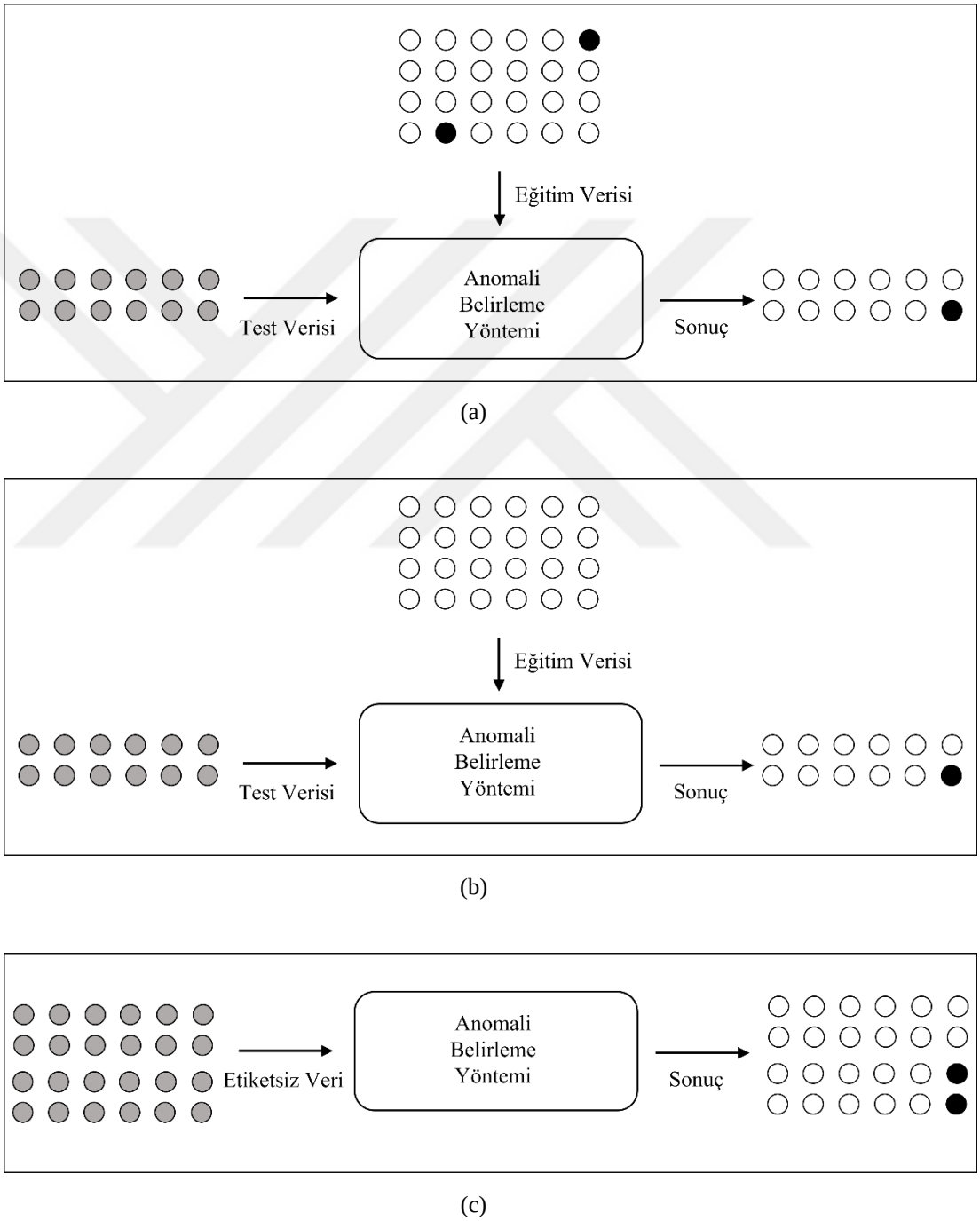
Tek bir gözlem olarak ele alındığında anomali olarak görülmeyen ancak birçok gözlemin bir araya gelerek oluşturdukları örüntü ile verinin geri kalan desenine göre farklılık gösteren gözlem gruplarına ise kolektif anomali denir (Chandola vd., 2009). Özellikle sağlık alanında kardiyovasküler rahatsızlıkların tespitinde yaygın olarak kullanılan "electrocardiogram" (EKG) kolektif anomaliye güzel bir örnektir. Şekil 2.4'te örnek bir EKG'de anomali olarak tespit edilen kısım işaretlenmiştir.



Şekil 2.4 Kolektif anomali örneği (Keogh, Lonardi ve Chiu, 2002).

2.3 Anomali Belirleme Yaklaşımları

Anomali belirleme yaklaşımları, veride anomali durumunu niteleyen bir etiket bulunup bulunmamasına göre denetimli anomali belirleme, yarı denetimli anomali belirleme ve denetimsiz anomali belirleme olarak üç ana başlıkta incelenebilir.



Şekil 2.5 Anomali belirleme yöntemleri (Goldstein ve Uchida, 2016)

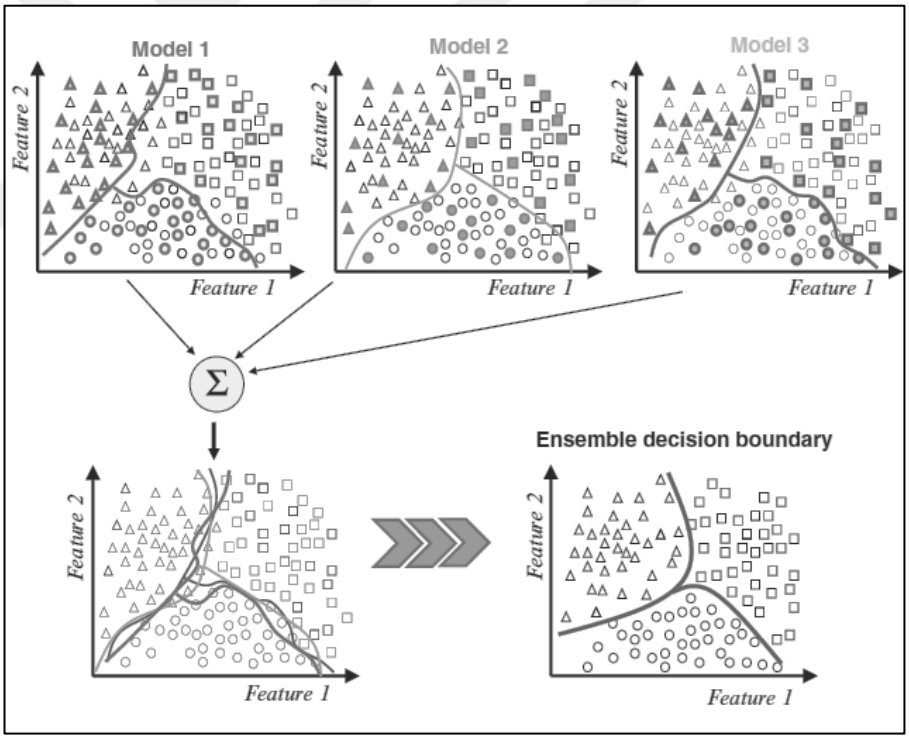
Model oluşturmak için hem olağan hem de anomali gözlemlerini içeren etiketli eğitim verileri gerektiren yaklaşıma denetimli anomali belirleme denir. Denetimli anomali belirleme yöntemleri dengesiz sınıflama problemi olarak da tanımlanabilir. Naive Bayes, K-En Yakın Komşu, Destek Vektör Makineleri, Karar Ağaçları, Yapay Sinir Ağları denetimli anomali belirlemede kullanılabilen yaygın yöntemlerdir (Zhou, 2012). Şekil 2.5 (a)'da hem normal hem de anomali gözlemlerin bulunduğu veri ile eğitilen denetimli anomali belirleme görselleştirilmiştir.

Eğitim sırasında sadece normal gözlemlerle eğitilen, normal harici bir gözlem ile karşılaşıldığında bu durumu anomali olarak belirleyen yaklaşıma yarı denetimli anomali belirleme denir. Bu yaklaşım, özellikle endüstride makine arızası tespitinde ve gözetleme sistemlerinde kullanılmaktadır. Literatürde yenilik belirleme (novelty detection) veya tek sınıflı sınıflandırma (one-class classification) olarak ele alınmaktadır (Aggarwal, 2017). Şekil 2.5 (b)'de yarı denetimli anomali belirleme görselleştirilmiştir.

Herhangi bir veri etiketi bulunmaksızın verinin kendi iç özellikleri ile, benzer yapıdaki gözlemlerin incelemesi yapılarak anomalilerin belirlendiği yaklaşıma denetimsiz anomali belirleme denir. Kümeleme, histogram ve komşuluk tabanlı yöntemler denetimsiz anomali belirlemede yaygın olarak kullanılır (Aggarwal, 2017).

BÖLÜM ÜÇ
TOPLULUK ÖĞRENMESİ İLE ANOMALİ BELİRLEME

Yaşamda, karşılaşılan meselelerle ilgili bir tercih yapılmak istendiğinde veya önemli bir karar verileceği zaman genelde o konu hakkında bilgisine tecrübesine güvenilen birden çok kişiden fikir almak yaygın rastlanılan bir tutumdur. Örneğin, cerrahi bir operasyon gerektiren bir tedavi kararını vermek için birden çok hekimin görüşünün alınması istenebilir. Çoğunluğun vereceği kararın tek bir kişinin görüşünden daha isabetli olacağı düşünülür ve çoğunluğun kararına daha çok güvenilir. Ancak, karar verirken herkesin düşüncesine de aynı oranda değer verilmeyebilir. Günlük hayattaki bu yaklaşımın makine öğrenmesindeki karşılığı topluluk öğrenmesidir (Sagi ve Rokach, 2018).



Şekil 3.1 Topluluk öğrenmesi (Polikar, 2012)

Tek bir model kurmak yerine birçok modelin verdiği sonucu birlikte değerlendirerek karar verme fikri topluluk öğrenmesi yöntemleri başlığı altında incelenmektedir. Topluluk öğrenmesi ile birden fazla algoritmanın bir arada

kullanılması, bir algoritmanın farklı değerler ile çalıştırılması sonucu elde edilen sonuçların birlikte değerlendirilmesi ya da verinin alt kümelerinden elde edilen sonuçların bir arada kullanılmasıyla daha isabetli tahminleme yapılması amaçlanmaktadır. Bu mantık ile daha güçlü bir genelleme kabiliyeti amaçlanmaktadır (Fernandes ve Carvalho, 2019).

Şekil 3.1'de farklı modellerden elde edilen tahminleme sonuçlarının bir arada değerlendirilmesiyle sınırların daha isabetli belirlenmesi görselleştirilmiştir. Birden fazla yöntemi veya bir yöntemin varyasyonlarından elde edilen sonuçları birlikte kullanma fikri hata varyansını ve aşırı uyum sorununu azaltacağı gibi farklı yöntemlerin zayıflıklarını telafi etme yönüyle de isabet oranı daha yüksek tahmin yapmaya katkı sağlayacaktır (Chawla ve Sylvester, 2007).

Bir topluluk öğrenmesi modelinin başarılı sayılabilmesi için kendisini oluşturan tüm modellerden daha iyi sonuç vermesi beklenir. Başarılı bir topluluk öğrenmesi modeli oluşturabilmek için topluluğu oluşturan her modelin diğeriyle tamamen aynı tahminleri yapmaması; farklı hatalar yapması gerekir. Bu durum modeller arasında korelasyon olmamasıyla sağlanabilir (Polikar, 2012).

Topluluk öğrenmesi yaklaşımı, anomali belirleme problemlerinde de kullanılmaktadır. Topluluk öğrenmesi yaklaşımı ile anomali belirleme üç başlık altında incelenebilir.

Birincisi; sınıf etiketi bulunan veride sınıflar arasında oransal farkın oldukça yüksek olduğu durumlarda topluluk öğrenmesi yöntemleriyle sınıflama problemi bir anomali belirleme olarak düşünülebilir. Bu şekilde yapılan anomali belirleme, topluluk öğrenmesi ile denetimli anomali belirleme olarak adlandırılır. İkincisi, gözlemlerden çoğunun benzer özellikler taşıdığı varsayılan etiketi bulunmayan veride topluluk öğrenmesi yöntemleriyle analiz yapılarak küçük grupta yer alan gözlemlerin anomali olarak nitelendirilmesi durumudur. Bu durumda anomali belirleme, topluluk öğrenmesi ile denetimsiz anomali belirleme olarak adlandırılır. Üçüncüsü, sadece normal gözlemler bulunan veride topluluk öğrenmesi yöntemleriyle oluşturulan

modelin farklı nitelikte az sayıda gözlem ile karşılaştığında bunu anomali olarak ayırt ettiği durumdur. Bu durum topluluk öğrenmesi ile yarı denetimli anomali belirleme olarak adlandırılır (Aggarval ve Sathe, 2017).

Sınıfları arasında oransal olarak dengesizlik bulunan veride azınlık sınıfına ait gözlemlerin anomali olarak sınıflandırılması bir denetimli anomali belirleme problemi olarak ele alınabilir. Topluluk öğrenmesi yaklaşımı ile denetimli anomali belirleme yöntemlerinden önce sınıflandırma yapılacak veride dengesizlik olduğu durumlarda sınıflandırma algoritmalarının çalışmasını etkileyen sınıf dengesizliği problemi incelenmelidir. Takip eden bölümlerde önce dengesiz sınıf problemine değinilmiş ve sonrasında topluluk öğrenmesi yöntemleri ile denetimli ve denetimsiz anomali belirleme yöntemleri açıklanmıştır. Bölüm sonunda ise model değerlendirmede kullanılan performans ölçütlerine yer verilmiştir.

3.1 Sınıf Dengesizliği Problemi

Gerçek verilerin sınıflara ait gözlem sayıları genellikle aynı olmaz. Bu fark oransal olarak arttığında sınıflar arası dengesizlik ortaya çıkar. Sınıf dengesizliği problemi, bir veride bir sınıfa ait gözlemlerin diğer sınıfa göre daha az oranla bulunması durumudur. İki sınıflı ve daha çok sınıflı veriler için geçerli olan bu husus sınıflandırma yöntemlerinin beklenen performansı gösterememesine neden olmaktadır (He ve Garcia, 2009). Dengesiz veride azınlık sınıfının diğer sınıfa/sınıflara oranının tanımlama için kabul edilmiş bir sınırı olmamakla beraber Krawczyk (2016) bu oranın 1/4 ile 1/100 arasında olabileceğini, bazı gerçek uygulama senaryolarında 1/1000, 1/5000 gibi aşırı dengesizlik oranlarının da görülebildiğini belirtmiştir.

Makine öğrenmesi algoritmalarının başarısı kullanılan algoritmanın, parametre değerlerinin yanı sıra verinin yapısına da bağlıdır. Sınıflar arasında dengesizlik olan verilerde, sınıflardan birisi daha çok temsil edilirken diğeri çok daha az temsil edilir. Makine öğrenmesi algoritmaları baskın bulunan sınıf üzerinde yüksek doğruluk değerine ulaşma eğilimindedirler. Bu durum sınıflandırmada yanlılık olarak tanımlanır (Chawla, Bowyer, Hall ve Kegelmeyer, 2002).

Sınıf dengesizliği sadece sınıflar arası değil; aynı sınıfa dâhil gözlemler arasında da olabilir (He ve Garcia, 2009). Bu duruma sınıf içi dengesizliği denilmekte olup daha çok verinin dağılımı ile ilgilidir. Bu konu çalışmanın kapsamı dışında tutularak, sınıflar arası dengesizlik konusu anomali belirleme kapsamında ele alınmıştır.

Anomali belirleme yaklaşımları genellikle anomalilerin nadir karşılaşılan durumlar/gözlemler olduğu varsayımını yapar. Bu varsayıma göre gözlemleri olağan ve anomali olarak sınıflandırmak dengesiz sınıflama problemi olarak ele alınabilir. Bu problemde, sınıf dengesizliği bulunan veride azınlık sınıfına ait örnekleri doğru sınıflandırmak daha önemlidir. Çünkü ilgilenilen sınıf azınlık olandır. Dengesiz veride az sayıda bulunan anomalileri doğru sınıflandırabilmek için çeşitli yöntemler önerilmiştir. Bu yaklaşımlar veri seviyesinde ve algoritma seviyesindeki çözümler olarak iki başlık altında incelenmektedir (Haixiang vd., 2017).

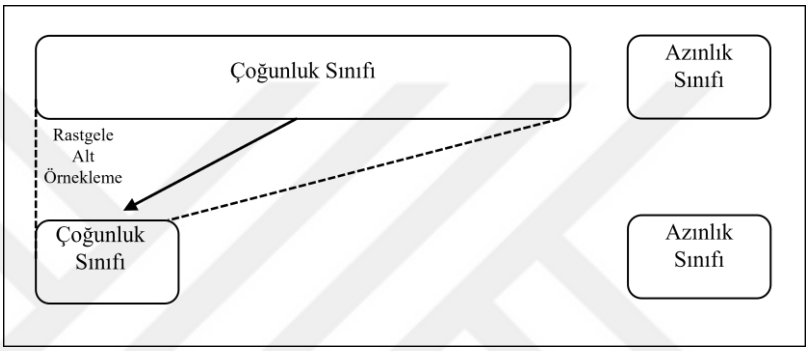
Veri seviyesindeki çözümler yeniden örnekleme ile yanlılık etkisini azaltmayı amaçlar. Algoritma seviyesinde çözümler ise makine öğrenmesi algoritmalarına anomali değerleri yakalamaları için basamakları arasında ilave tedbirler alınarak geliştirilen yeni algoritmalardır. Bu iki yaklaşımın bir arada kullanıldığı hibrit yaklaşımlar da mevcuttur (Krawczyk, 2016).

3.1.1 Veri Seviyesinde Çözümler

Veri seviyesinde çözümlerde öğrenme, algoritmadan bağımsız veri sınıflarının dağılımının dengelenmesi esasına dayanır. Öğrenme yeniden örnekleme yöntemleri uygulanmış veri üzerinden gerçekleştirilir. Yeniden örnekleme çözümleri, sınıflandırma algoritmalarından bağımsız yöntemler olduğu için tüm algoritmalar için kullanılabilecek genel çözümlerdir (Fernandez vd., 2018).

Yeniden örnekleme, sınıf dağılımındaki dengesizliği giderebilmek için veri setinde azınlık sınıfı örneklerin artırılması, çoğunluk sınıfı örneklerin azaltılması veya azınlık sınıfı temsil eden yeni örneklerin sentetik olarak üretilmesi ile yapılır (Das, Krishnan ve Cook, 2014).

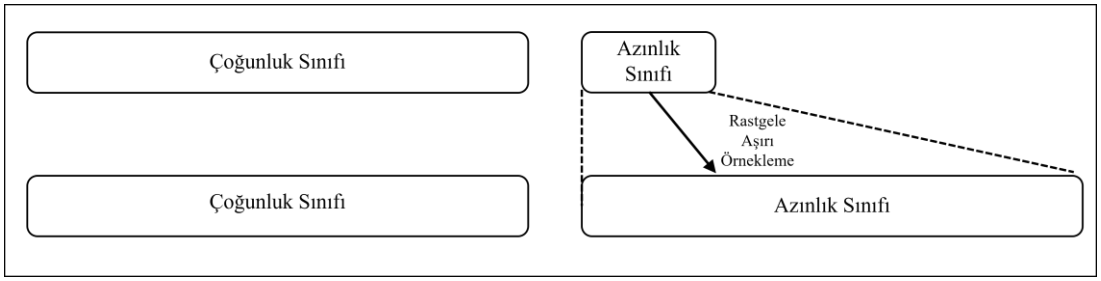
Rastgele alt örnekleme (RUS) ile eğitim setinden çoğunluk sınıf örnekleri rastgele kaldırılır. Böylece sınıflandırma esnasında iki sınıfın da yakın oranda temsil edilmesi sağlanmış olur. (Baesens, Vlasselaer ve Verbeke, 2015). Ancak çoğunluk sınıftaki elamanlar kaldırılırken tekrar eden gözlemlerin kaldırılması amaçlanması gerekirken rastgele veri azaltma işlemi yapılması sınıflandırmaya ait yararlı bilginin kaybolmasına da neden olabilir (He ve Garcia, 2009). Bu yöntem basitliği ve eğitim süresinin azalmasına olan katkısı nedeniyle sıklıkla kullanılır (Drummond ve Holte, 2003). Şekil 3.2'de rastgele alt örnekleme işlemi gösterilmiştir.



Şekil 3.2 Rastgele alt örnekleme

Rastgele aşırı örnekleme (ROS) yöntemi, rastgele alt örnekleme yönteminin tersidir. Yani azınlık sınıfında bulunan örneklerin rastgele seçilerek çoğaltılması esasına dayanır. Böylece eğitim verisinde azınlıkta bulunan sınıf ile çoğunluk sınıfının örnek sayısı birbirine yaklaşmış olur. (Baesens vd., 2015).

Bu yöntem, basit olması ve azınlık sınıfının öğrenilmesini olumlu yönde etkilemesinin yanında aşırı uyum (overfitting) problemine de neden olabilmektedir (Batista vd., 2015). Diğer dezavantajı ise eğitim kümesinin boyutunun artması nedeniyle eğitim süresinin uzamasına sebep olmasıdır. (Holte ve Drummond, 2003). Şekil 3.3'te rastgele aşırı örnekleme işlemi gösterilmiştir.



Şekil 3.3 Rastgele aşırı örnekleme

Literatürde, azınlık sınıfı örneklerin rastgele çoğaltılması yerine modelin eğitildiği veride azınlık sınıfa ait verinin yapay olarak çoğaltılmasına dayalı yöntemler önerilmiştir. Bu yöntemler, birbirine yakın iki azınlık sınıfı gözlemi arasında o sınıfı iyi temsil edecek yapay veri üretme mantığına dayanır ve ilk olarak Chawla vd. (2002) tarafından Synthetic Minority Over-Sampling Technique (SMOTE) ismiyle önerilmiştir.

1: Azınlık sınıfının her bir gözlemi için k en yakın komşusu bulunur.

2: Azınlık sınıfındaki her bir gözlem ile k en yakın komşularının ilgili gözlem ile farkı alınır.

3: 0 ve 1 arasında rastgele bir sayı (α) belirlenir ve ikinci aşamada hesaplanan fark ile bu rastgele sayı çarpılır.

4: $x_{yeni} = x_i + (x_j - x_i) * \alpha$ eşitliği kullanılarak yeni yapay gözlem oluşturulur.

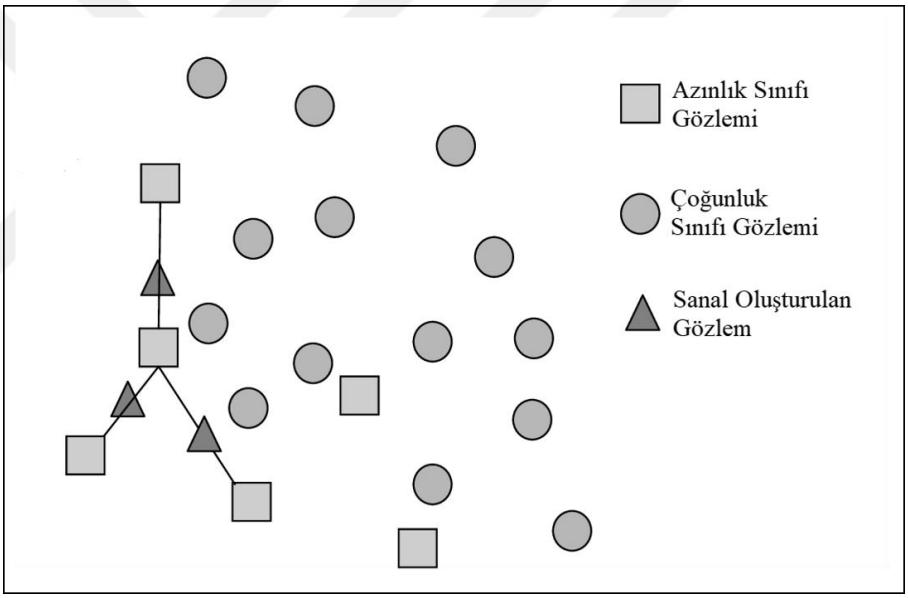
5: İlk 4 adım yeterli sentetik gözlem elde edilene kadar devam ettirilir.

Şekil 3.4 SMOTE algoritması

SMOTE yöntemi rastgele aşırı örnekleme ve rastgele alt örnekleme yöntemlerinin dezavantajlarını taşımadan avantajlarını birleştirdiği için sonraki yıllarda literatüre çok

sayıda varyasyonu önerilmiştir (Fernandez vd., 2018). Rastgele aşırı örnekleme yönteminin aksine azınlık sınıfı örnekleri rastgele çoğaltmak yerine azınlık sınıfı örneklerinden o sınıfın özelliklerine uygun, azınlık sınıfını iyi temsil edecek yeni yapay örnekler oluşturulur. Böylece, rastgele aşırı örnekleme yönteminin temel sorunu olan aşırı uyum sorununun önüne geçilir (Estabrooks, Jo ve Japkowicz, 2004). SMOTE yöntemi temel olarak Şekil 3.4'teki gibi çalışır (Chawla vd., 2002). Şekil 3.5'te ise SMOTE yöntemi ile azınlık sınıfına yapay örnek üretilmesi gösterilmiştir.

Orijinal SMOTE yöntemi, çoğunluk sınıfını dikkate almadan işlem yaptığından, daha sonra çoğunluk sınıfını da dikkate alarak sentetik örnek oluşturan farklı yöntemler önerilmiştir.



Şekil 3.5 SMOTE yöntemi ile azınlık sınıfına yapay örnek üretilmesi (Sun vd., 2018)

Azınlık sınıfı örneklerini üretirken sınıflandırması daha zor olduğu düşünülen örnekleri dikkate alınarak yapay gözlemler oluşturan Adaptive Synthetic Sampling (ADASYN) en çok kullanılan SMOTE yöntemlerinden biridir (He, Bai, ve Garcia, 2008). ADASYN algoritmasında sentetik veriler veri yoğunluğuna göre oluşturulur. Yoğunluğun düşük olduğu verilere odaklanılır. Bu tez çalışmasının uygulama bölümünde yeniden örnekleme metodu olarak rastgele aşırı örnekleme ile birlikte ADASYN tercih edilmiştir.

3.1.2 Algoritma Seviyesinde Çözümler

Algoritma seviyesinde çözümlerde örnek olarak Maliyet Duyarlı Öğrenme (Cost Sensitive Learning) gösterilebilir (Hoens ve Chawla, 2013). Maliyet Duyarlı Öğrenmede azınlık sınıfındaki bir örneğin yanlış sınıflandırılması, çoğunluk sınıftaki bir örneğin yanlış sınıflandırılmasından daha fazla öneme sahiptir. Bu mantıkla yanlış sınıflandırılan örneklere sınıfına göre ceza puanı (penalties) atanır ve yanılgı matrisine benzer maliyet matrisi oluşturulur (Zhou ve Liu, 2005). Maliyet matrisi ve maliyet analizi Bölüm 3.4.8'de açıklanmıştır.

3.2 Denetimli Topluluk Öğrenmesi Yöntemleri

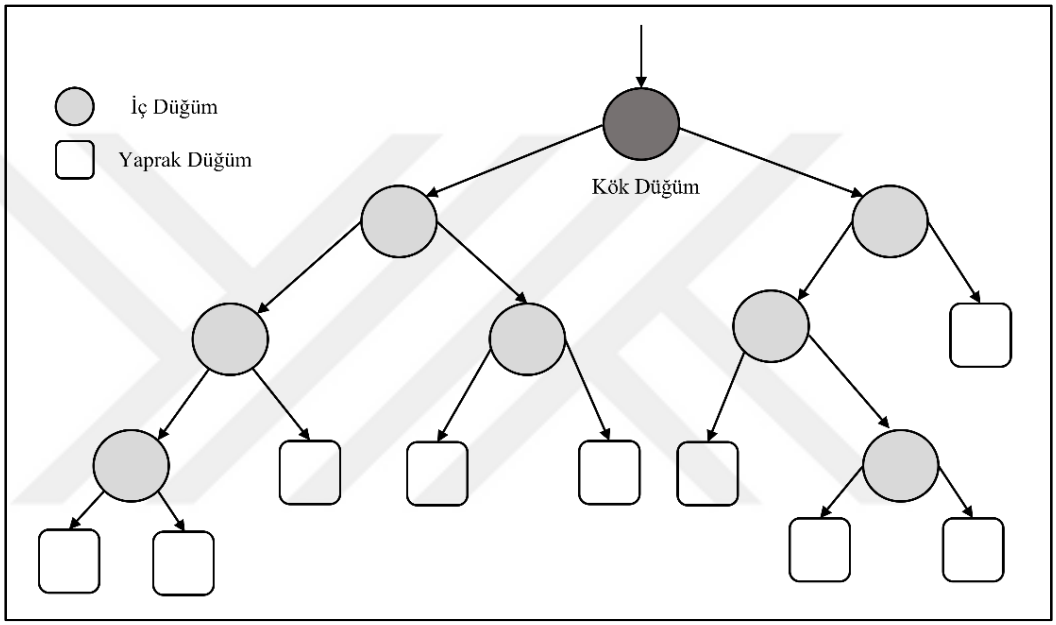
Denetimli topluluk öğrenmesi yöntemleri örneklem seçimi, birleştirme işlem adımları ve değerlendirme tekniğine göre Bagging, Boosting, Oylama ve Stacking başlıkları altında incelenir. Bazı Türkçe çalışmalarda Bagging için Torbalama, Boosting için Artırma veya Yükseltme, Stacking için Yığınlama veya Yığma isimlerinin kullanıldığı görülmüştür. Ancak bu çalışmada orijinal kullanımları tercih edilmiştir.

3.2.1 Bagging

Breiman (1996) tarafından önerilmiştir. **B**ootstrap **AGG**regat**ING** kelimelerinin kısaltılarak birleştirilmesiyle oluşturulmuş bir ifadedir (Polikar, 2012). Bootstrap yöntemi ise Efron (1992) tarafından önerilmiş bir yeniden örnekleme yaklaşımıdır. Özellikle bilgisayarla hesaplama imkanlarının gelişmesiyle istatistiksel çıkarsamadan model değerlendirmeye birçok farklı alanda Bootstrap yaygın olarak uygulanmaktadır. Kısaca, veriden tekrarlı rastgele aynı genişlikte örneklemler elde etmek olarak tanımlanabilir. Aggregating ise rassal örneklemlerden elde edilen çıkarsamaların bir bütün olarak bir arada değerlendirilmesidir.

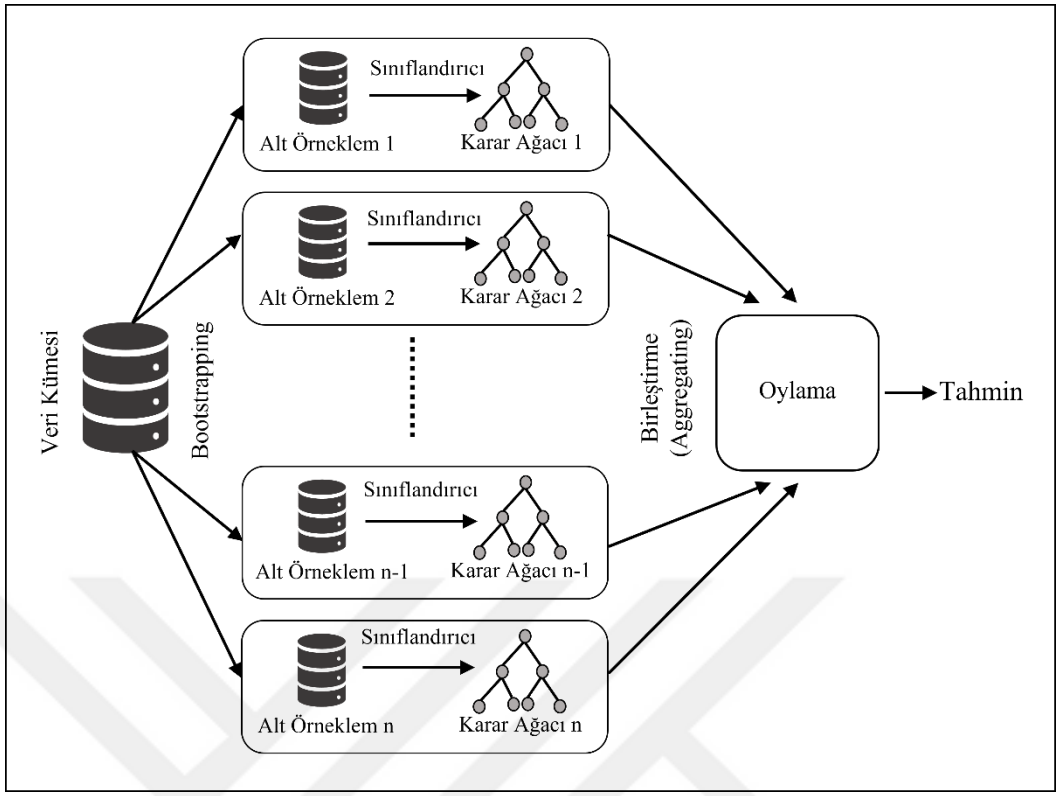
Bagging yöntemi ve bir sonraki başlıkta anlatılan Boosting yöntemleri karar ağaçlarını kullanırlar. Sınıflandırma ağaçları (CT) makine öğrenmesinde kullanılan

modelleme yaklaşımlarından biridir (Sutton, 2005). Sınıflandırma ağacı öğrenme esnasında, öğrendiği bilgiyi bir ağaç üzerinde işler. Kök düğüm karar ağacının ilk elemanıdır. Bu düğümün altında bulunan düğümler iç düğüm veya bir alt düğümü olmayan karar değerini taşıyan yaprak düğüm olabilir. Şekil 3.6'da bir sınıflandırma ağacının yapısı görselleştirilmiştir. Literatürde birçok karar ağacı algoritması bulunmaktadır. En iyi bilinen karar ağacı algoritmaları ID3 ve onun geliştirilmiş versiyonu C4.5 ve CART algoritmalarıdır (Kuhn ve Johnson, 2013).



Şekil 3.6 Bir sınıflandırma ağacı modeli örneği

Karar ağacı ile sınıflandırmada tek bir ağaç ile tahmin yapılmaktadır. Bu da yanlılığa sebep olmaktadır. Bagging yönteminde ise tek bir karar ağacı kullanmak yerine orijinal veriden rastgele tekrarlı örneklemler ile yeni karar ağaçları elde edilir (Aggarval ve Sathe, 2017). Böylece birbirine paralel öğrenen çok sayıda karar ağacı topluluğu elde edilmiş olur. Her bir karar ağacı farklı bir örneklemden öğrenmiş olur. Her bir karar ağacının verdiği tahmin değeri dikkate alınarak bir oylama ile tahmin gerçekleştirilir. Böylece, karar ağacı ile sınıflandırma yönteminde ortaya çıkan yanlılık ve aşırı uyum sorununun önüne rassallık sağlanarak geçilmiş olur, varyans azalır ve tahmin başarısında artış sağlanmış olur. Şekil 3.7'de Bagging yönteminin temel çalışma yapısı gösterilmiştir.

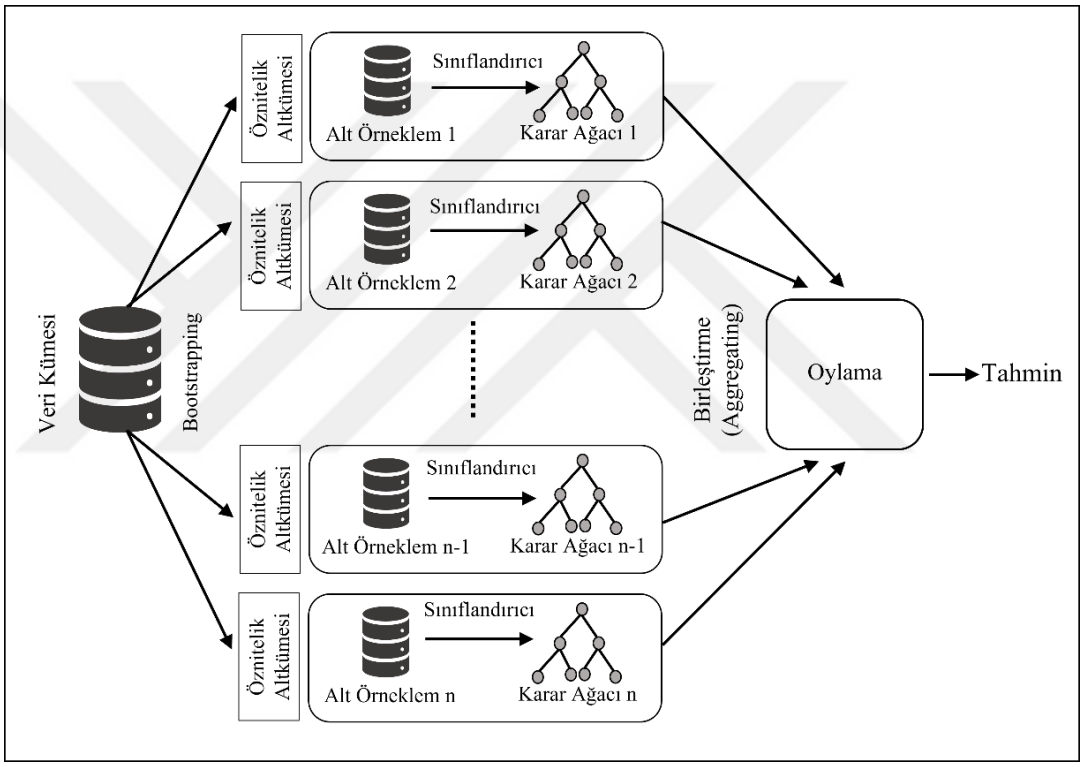


Şekil 3.7 Bagging yönteminin çalışma yapısı

Bagging yönteminde kullanılacak karar ağacı sayısına karar verilmesi gerekir. Model performansının artmadığı görülene kadar ağaç sayısının artırılması tercih edilir. Karar verilmesi gereken diğer parametre ise örneklem büyüklüğüdür. Orijinal veri seti ile aynı boyutta örneklem seçilebileceği gibi daha küçük örneklemler de oluşturulabilir. Örneklemler iadeli oluşturulduğu için bazı gözlemler bazı örneklemlerde birden çok defa bulunabileceği bazıları gözlemler bazı örneklemlerde hiç bulunmayabilir.

Bagging yönteminin literatürde önerilmiş farklı varyasyonları vardır. Veriye ait özniteliklerin rastgele alt kümelerini almaya dayalı Random Subspaces (Ho, 1998), tüm veriyi kullanarak ağaçlar oluşturmaya dayalı Extra Trees (Geurts, Ernst ve Wehenkel, 2006) ve Bagging algoritması ile Random Subspaces metodundan esinlenerek oluşturulan Random Forest (Breiman, 2001) algoritması bunlardan bazılarıdır.

Random Forest (RF) algoritmasının Bagging algoritmasından önemli farkı, verinin özniteliklerinin yani bağımsız değişkenlerin de rastgele alt kümesinin alınmasıyla eğitim yapılmasıdır. Yani gözlemlerde yeniden örnekleme yapılmasının yanında değişkenlerin de rastgele alt kümeleri oluşturularak öznitelik altkümeleri ile karar ağaçları oluşturulur. Bagging yönteminde olduğu gibi birden çok sayıda karar ağacı birlikte değerlendirilerek karar verildiği için aşırı uyum sorununun önüne geçilmiş olur (Breiman, 2001). Şekil 3.8'de Random Forest yönteminin temel çalışma yapısı gösterilmiştir.



Şekil 3.8 Random Forest algoritmasının çalışma yapısı

Breiman (2001) seçilecek değişkenlerin adedinin belirlenmesinde N değişkenli bir veri seti için $\log_2(N + 1)$ adet değişken olmasını önerirken Hastie, Tibshirani ve Friedman (2009) $\sqrt{N}$ değerini önermişlerdir. Tez çalışmasının uygulama bölümünde Bagging yöntemlerinden Random Forest algoritmasının kullanımı tercih edilmiştir. Öznitelik altkümesi sayısı belirlenirken $\sqrt{N}$ değerinin bir üst tam sayıya yuvarlanmasıyla elde edilen değer kullanılmıştır.

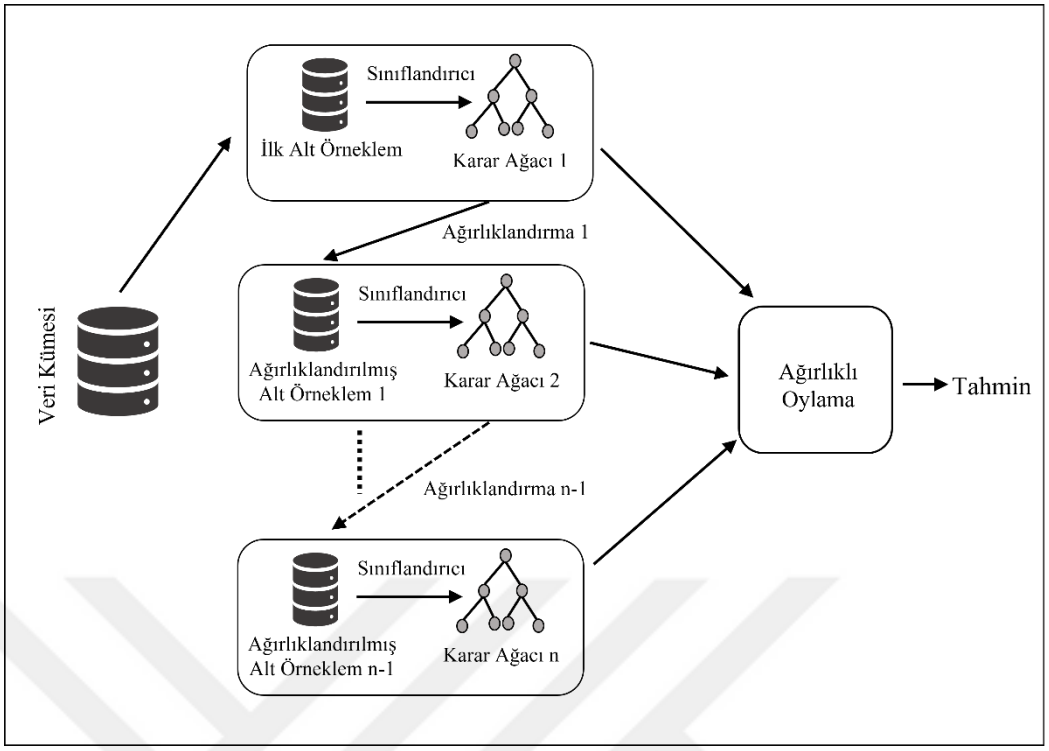
3.2.2 Boosting

Schapire (1990), Kearns ve Valiant (1994)'ın zayıf öğrenicilerin birlikte kullanılmasıyla hata oranının azaltılmasından bahsettiği öğrenme teorisi ile ilgili çalışmalarından etkilenerek Boosting yöntemini önermiştir. Bagging yönteminden farklı olarak ardışık ilerlemeye dayalı bir yöntemdir. Topluluk öğrenmesi yöntemleri arasında oldukça popüler hale gelmiştir. Zayıf öğrenicileri bir arada kullanarak daha iyi bir öğrenici elde etmeyi, önceki eğitim adımlarında öğrendiği bilgiyi sonraki adıma taşımayı hedefler (Schapire, 1990).

Boosting yönteminde sınıflandırıcılar önceki aşamadaki eğitim sonucunu dikkate alarak öğrenmeye devam ederler. Böylece önceki aşamada yanlış öğrenilen örnekler yeniden örnekleme yapılırken daha fazla seçilme sıklığına sahip olması hedeflenerek o gözlemlere ağırlık verilmesi sağlanır (Schapire, 1990).

Başlangıçta her örneğe eşit ağırlık verilerek önceki eğitim adımlarında yanlış öğrenilen örnekler bir sonraki eğitim adımında daha yüksek ağırlıklar verilmesiyle öğrenme gerçekleştirilir. Böylece varyans ve yanlılık azaltılır (Aggarval ve Sathe, 2017).

Eğitim aşamasının sonunda her sınıflandırıcı ile oluşturulan modelin ağırlığı, eğitim seti üzerindeki doğruluk oranı ile orantılı belirlenerek oylama ya da ağırlıklı ortalama yöntemi ile birleştirme yapılır. Şekil 3.9'da Boosting yönteminin temel çalışma yapısı gösterilmiştir.



Şekil 3.9 Boosting yönteminin yapısı

Bagging yönteminde öğreniciler birbirinden bağımsız paralel olarak öğrenirken Boosting yönteminde öğreniciler birbirine sıralı bağlıdırlar. Bagging yöntemi aşırı uyum problemine oldukça dayanıklıyken Boosting yönteminde seçilen parametrelerin değerleri dikkatle belirlenmezse aşırı uyum sorununa sebep olabilmektedir. Boosting yönteminde, Bagging yönteminde tercih edilen alt örneklem kullanımı tercih edilmez (Aggarval ve Sathe, 2017).

En bilinen Boosting algoritmaları Adaptive Boosting (AdaBoost) (Freund ve Schapire, 1997) algoritması, Gradient Boosting (GBM) (Friedman, 2001) algoritması ve Extreme Gradient Boosting (XGBoost) (Chen ve Guestrin, 2016) algoritmalarıdır.

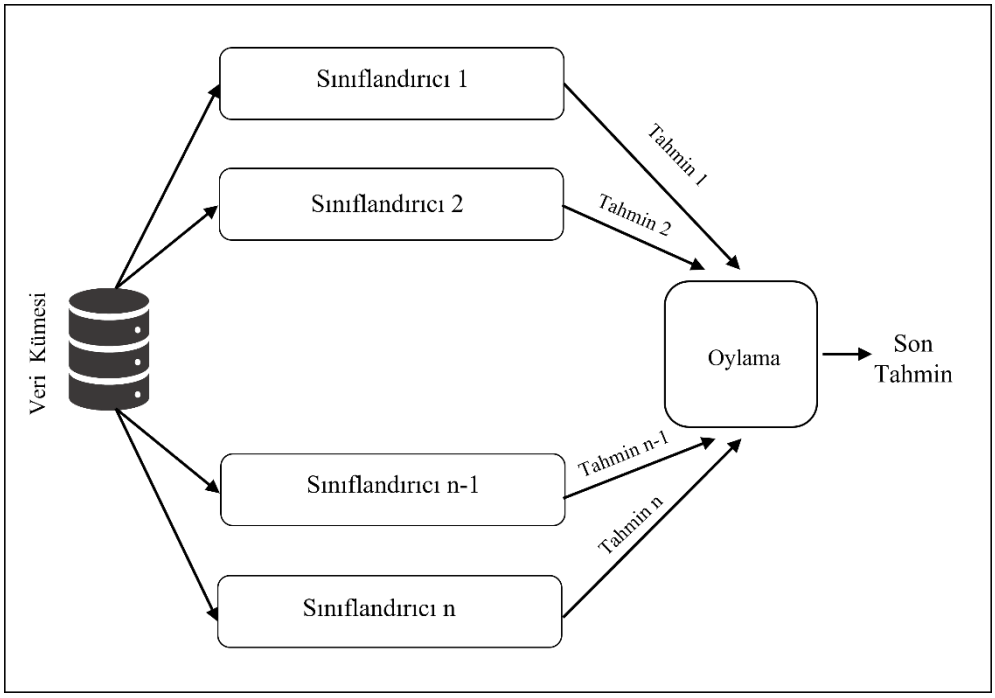
Adaptive Boosting (AdaBoost) algoritması verinin alt kümelerini eğitmek yerine ilerlemeli olarak farklı ağırlıklarla verinin eğitilmesi esasına dayanır. Her adımda önceki adımda alınan hatalara odaklanarak hatanın azaltılması amaçlanır (Freund ve Schapire, 1997).

Gradient Boosting (GBM) algoritması ise hata fonksiyonu sonucunun küçültülmesini hedefler. Gradient Descent yöntemiyle her yinelemeli adımda bir önceki adım hata fonksiyonunun minimuma düşürülmesi sağlanır (Friedman, 2001).

Extreme Gradient Boosting (XGBoost), Chen ve Guestrin (2016) tarafından Gradient Boosting algoritmasının geliştirilmesiyle literatüre kazandırılmış ve son yıllarda oldukça popüler olmuş bir Boosting algoritmasıdır. Dağıtık sistemlerde çalışabilmesi ve paralel işleme kabiliyeti sayesinde daha hızlı ve esnek olmasının yanında, daha fazla değiştirilebilir parametreye sahip olması nedeniyle parametre seçiminin iyi yapılması gerekmektedir (Chen ve Guestrin, 2016). Tez çalışmasının uygulama bölümünde Boosting yöntemlerinden XGBoost algoritmasının kullanımı tercih edilmiştir.

3.2.3 Oylama (Voting)

Farklı sınıflandırma algoritmalarının verdiği sonuçların her bir gözlem için oylama sonucuna göre sınıflandırılmasına dayanır.

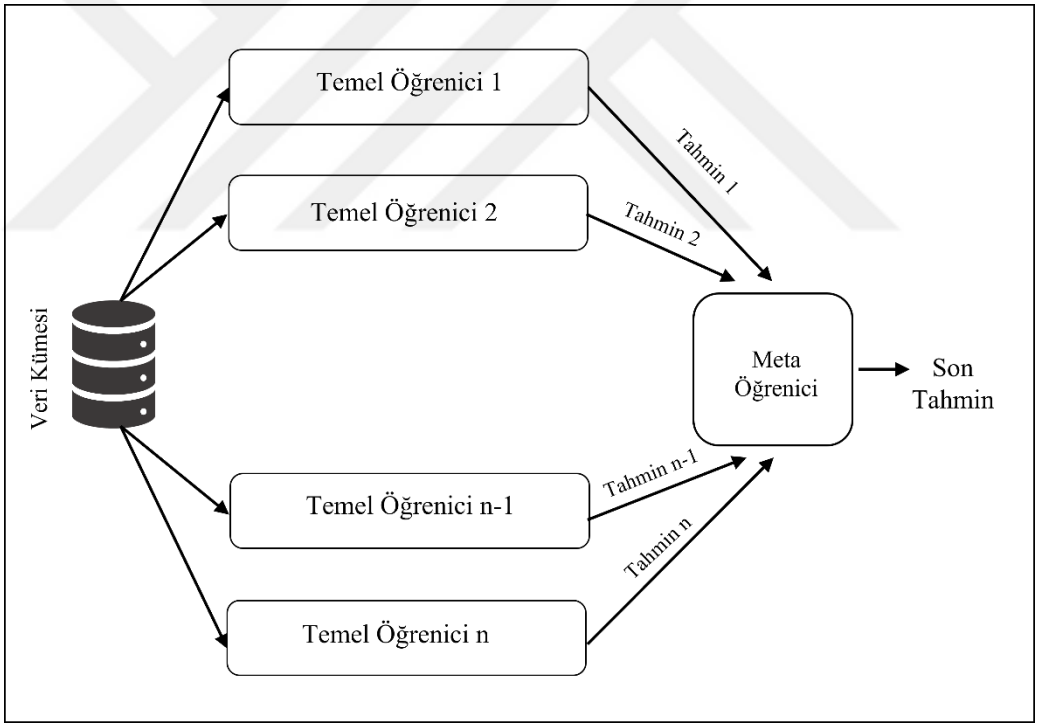


Şekil 3.10 Oylama yönteminin yapısı

Oylama, çoğunluk oyu veya model performansına göre her sınıflandırıcının farklı ağırlığa sahip olduğu ağırlıklı oylama olabilir (Polikar, 2012). Şekil 3.10'da Oylama yönteminin temel çalışma yapısı gösterilmiştir.

3.2.4 Stacking

Uygulamasının zorluğundan dolayı Bagging ve Boosting yöntemlerime göre daha az uygulanan bir denetimli topluluk öğrenmesi yöntemidir. İlk seviye modellerin sonuçlarını değerlendiren ikinci seviye model ile tahminleme yapılması esasına dayanır. Yani öğrenicilerden öğrenen bir model iki aşamalı tahminleme gerçekleştirilir. Bu modele meta-learner denir (Zhou, 2012). Şekil 3.11'de Stacking yönteminin temel çalışma yapısı gösterilmiştir.



Şekil 3.11 Stacking yönteminin yapısı

Uygulanmasının zor olma nedeni farklı yollardan öğrenen birbiri ile benzer tahmin hataları içermeyen öğrenicilerin bir araya getirilmesi gerekliliğidir. Ancak genellikle

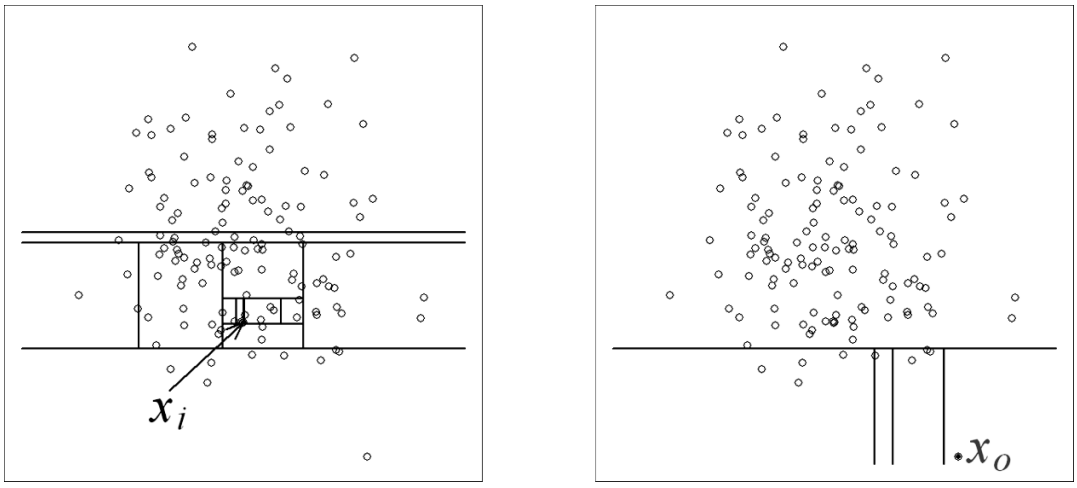
aynı veriden öğrenen öğreniciler birbirinden bağımsız olamamaktadır (Wolpert, 1992).

3.3 Denetimsiz Topluluk Öğrenmesi Yöntemleri

Denetimsiz öğrenmede eğitim verisinde sınıf etiketleri yoktur. Bu yüzden öğrenme yöntemleri veri içerisinde ortak noktalar veya kalıplar bulmaya çalışır. Bu ortaklıklara ve kalıplara göre gruplama yapılır. Böylece çoğunluk grubunda yer almayan gözlemler anomali olarak ayırt edilebilir. En bilinen denetimsiz öğrenme yöntemleri kümeleme ve birliktelik analizidir (Chandola, Banerjee ve Kumar, 2009). Topluluk öğrenmesi yöntemlerinden anomali belirleme amaçlı kullanılan en bilinen yöntem ise Isolation Forest yöntemidir (Aggarval ve Sathe, 2017).

3.3.1 Isolation Forest (iForest)

Anomali bir gözlemin normal bir gözlemden daha kolay izole edileceği hipotezi üzerinden geliştirilmiştir. İkili arama ağaçları topluluğu kullanarak anomali tespiti yapılması amacıyla önerilmiştir (Liu, Ting ve Zhou, 2008).



Şekil 3.12 iForest ile normal (sol) ve anomali gözlemin (sağ) izole edilmesi

Algoritmanın eğitim aşamasında Isolation Tree denilen ikili arama ağaçlarından bir topluluk oluşturulur. Oluşturulan her ağaçta yeni bir gözlem için kök düğümden yaprak düğüme kadar ilerleme yapılarak arama derinliğine bakılır. Arama derinliği ortalaması yüksek olan gözlemler anomali olduğu değerlendirilir. Çünkü anomali gözlemler normal gözlemlerden daha kolay izole edilir.

Şekil 3.12'de normal bir gözlem (sol) ile anomali gözlemin (sağ) örnek bir ağaçta izole edilmesi görselleştirilmiştir (Liu, Ting ve Zhou, 2008). Soldaki şekilde X_i noktasının izole edilmesi için 12 adım gerektiği; sağ görseldeki X_0 noktası için ise 4 adımın yeterli olduğu görülmektedir. Bu şekilde tüm ağaçlar için adım sayısı hesaplanarak ortalaması alınır ve ortalaması yüksek gözlemlerin anomali olduğu değerlendirilir.

3.4 Performans Değerlendirme

Bu tez çalışmasının uygulama bölümünde denetimli anomali belirleme yöntemleri kullanılmıştır. Denetimli anomali belirleme yöntemlerinin performans değerlendirme ölçütlerine bu bölümde yer verilmiştir. Oluşturulan modellerin ne kadar başarılı anomali tespiti yapabildiğini belirlemek için yanılgı matrisinden yararlanarak (confusion matrix), doğruluk (accuracy), belirleyicilik (specificity), duyarlılık (sensitivity), kesinlik (precision), AUC değeri, kesinlik- duyarlılık (PR) eğrisi ile bu eğri altında kalan alan değeri ile F Beta puanı kavramları incelenmiştir.

3.4.1 Yanılgı Matrisi (Confusion Matrix)

Sınıflandırma yöntemlerinin tahmin performanslarını ölçmek ve karşılaştırmak için gerçek durum ile tahmin edilen durumu karşılaştırmak için yanılgı matrisinden yararlanılır.

Yanılgı Matrisi		Gerçek Durum	
		Pozitif	Negatif
Tahmin	Pozitif	Doğru Pozitif	Yanlış Pozitif
	Negatif	Yanlış Negatif	Doğru Negatif

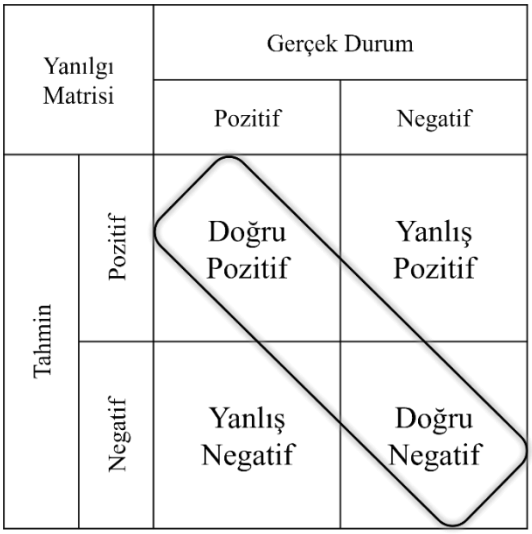
Şekil 3.13 Yanılgı Matrisi

Şekil 3.13'teki matrisin sütunları gerçek değerleri göstermektedir. Matrisin satırları ise sınıflandırma neticesinde bulunan tahmin değerlerini içermektedir.

Gerçekte pozitif olan bir gözlem pozitif olarak doğru sınıflandırıldığında Doğru Pozitif (True Positive) olarak adlandırılır. Gerçekte pozitif olan bir gözlem negatif olarak yanlış sınıflandırıldığında Yanlış Negatif (False Negative) olarak adlandırılır. Gerçekte negatif olan bir gözlem negatif olarak doğru sınıflandırıldığında Doğru Negatif (True Negative) olarak adlandırılır. Gerçekte negatif olan bir gözlem pozitif olarak yanlış sınıflandırıldığında Yanlış Pozitif (False Positive) olarak adlandırılır.

3.4.2 Doğruluk (Accuracy)

Sınıflandırma yapılırken en çok kullanılan ölçütlerdendir. Doğru tahmin sayısının toplam gözlem sayısına oranıdır. Şekil 3.14'te yanılgı matrisinde ilgilenilen kısım gösterilmiş ve Denklem 3.1'de formülü verilmiştir:



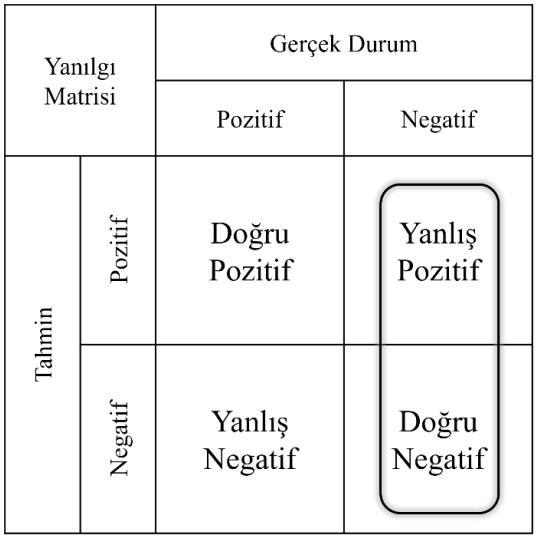
Şekil 3.14 Doğruluk

$$Doğruluk = \frac{\mathrm{DP} + \mathrm{DN}}{\mathrm{DP} + \mathrm{YP} + \mathrm{DN} + \mathrm{YN}} \tag{3.1}$$

Doğruluk, sınıfların dağılımının dengesiz olduğu durumlarda yeterli değildir (Krawczyk, 2016). Örneğin bilgisayar ağ trafiğinde tüm trafiğin %99'u normal akış, sadece %1'i saldırı trafiği olduğu bir senaryoda tüm ağ trafiğini normal trafik olarak sınıflandıran bir sınıflandırıcı %99 doğruluk değerine ulaşmış olur. Ancak esas tespit edilmesi istenen hiçbir ağ saldırı trafiği tespit edilememiş olur. Bu yüzden anomali belirlemede sadece doğruluk ölçütünün kullanılması uygun değildir (Menardi ve Torelli, 2014) ve bu tez çalışması kapsamında doğruluk değeri sonuçlarına yer verilmemiştir.

3.4.3 Belirleyicilik (Specificity, Selectivity)

Gerçekte negatif olan gözlemlerin ne kadarının negatif tahmin edildiğinin oranını verir. Belirleyicilik, sınıflandırmanın negatifleri doğru tahmin etme kabiliyetini gösterir. Şekil 3.15'te yanılgı matrisinde ilgilenilen kısım gösterilmiş ve Denklem 3.2'de formülü verilmiştir.

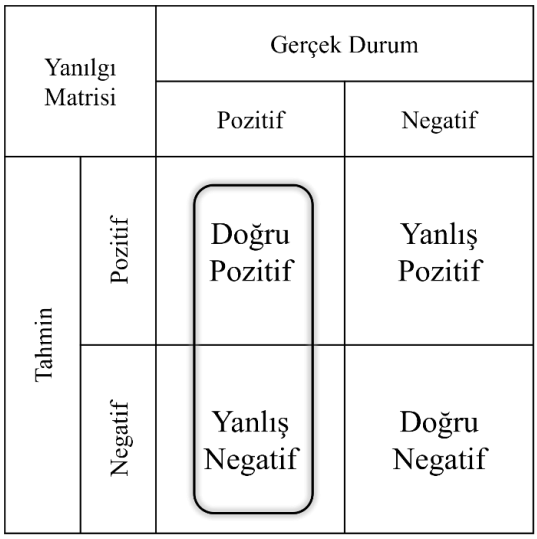


Şekil 3.15 Belirleyicilik (Doğru Negatif Oranı)

$$Belirleyicilik = \frac{\mathrm{DN}}{\mathrm{DN} + \mathrm{YP}} \tag{3.2}$$

3.4.4 Duyarlılık (Sensitivity, Recall)

Gerçekte pozitif olan bir gözlemin doğru tahmin edilme oranını verir. Duyarlılık sınıflandırmanın pozitifleri doğru tespit etme kabiliyetini gösterir.



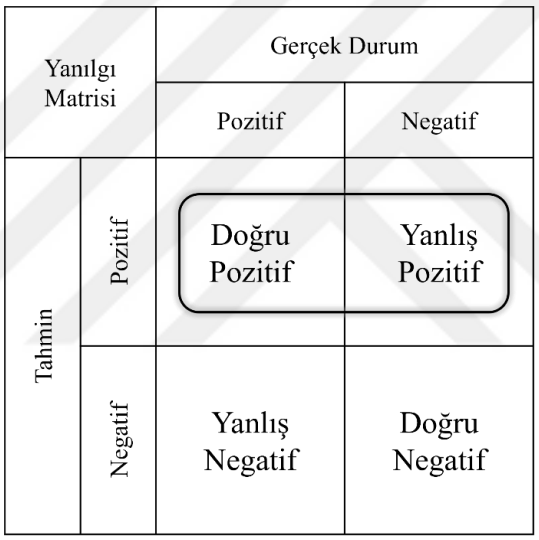
Şekil 3.16 Duyarlılık (Doğru Pozitif Oranı)

Şekil 3.16'da yanılgı matrisinde ilgilenilen kısım gösterilmiş ve Denklem 3.3'te formülü verilmiştir.

$$Duyarlılık = \frac{\mathrm{DP}}{\mathrm{DP} + \mathrm{YN}} \tag{3.3}$$

3.4.5 Kesinlik (Precision)

Pozitif olarak tahmin edilen gözlemlerin gerçekte ne kadarının pozitif olduğu ölçüsüdür.



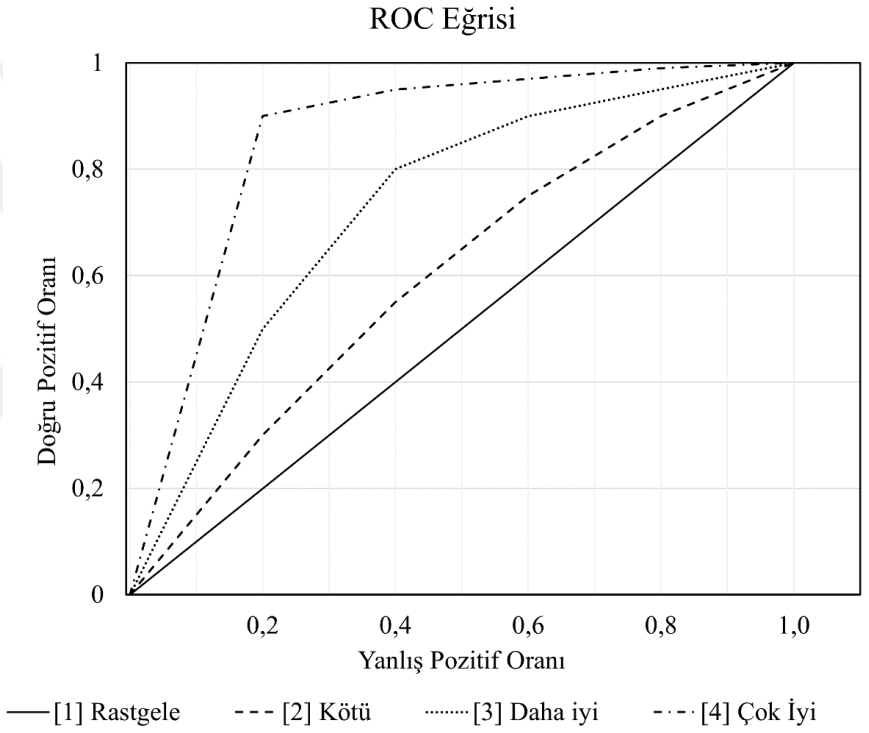
Şekil 3.17 Kesinlik

$$Kesinlik = \frac{\mathrm{DP}}{\mathrm{DP} + \mathrm{YP}} \tag{3.4}$$

Şekil 3.17'de yanılgı matrisinde ilgilenilen kısım gösterilmiş ve Denklem 3.4'te formülü verilmiştir.

3.4.6 ROC Eğrisi ve AUC Değeri

ROC eğrisi, duyarlılığın yani doğru pozitif değerinin y ekseninde, (1 – belirleyicilik) yani yanlış pozitif değerinin x ekseninde temsil edilmesi ile oluşturulmaktadır. Sınıflandırma yöntemleri duyarlılık ile belirleyicilik arasında dengeyi sağlamaya çalışır. Şekil 3.18'de gösterildiği gibi, eğri ne kadar dik bir şekilde yükselerek y ekseninde 1 değerine yaklaşırsa sınıflandırıcının o kadar iyi ayırt ediciliğe sahip olduğu söylenebilir.



Şekil 3.18 ROC eğrisi

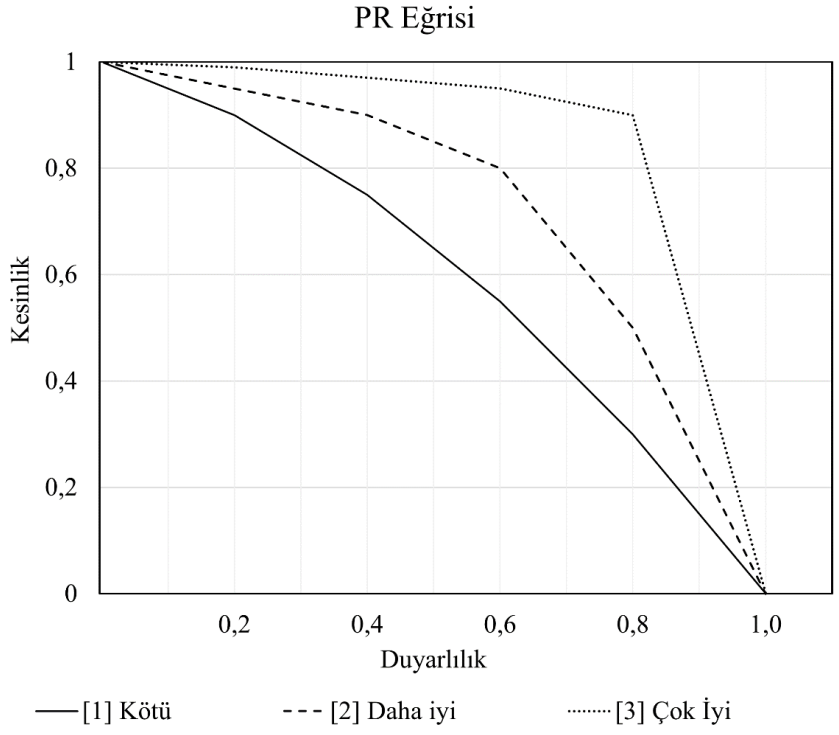
Farklı sınıflandırıcılar arasında, sınıflar arasında ayırma yapma gücünün karşılaştırmasında ROC eğrisinin altında kalan alan değerine göre inceleme yapılması çok tercih edilen bir yöntemdir (Davis ve Goadrich, 2006). Çünkü ROC eğrisinin dik bir şekilde yükselerek dikey eksende 1 değerine yaklaştığı durumda sınıflandırıcının daha iyi ayırt ediciliğe sahip olduğu söylenebilir. Bu durumda daha dik yükselen

sınıflandırıcının eğri altında kalan alan değeri daha büyük olacaktır. Bu değere eğri altında kalan alan (AUC) puanı denilmektedir. Sıfır ile bir arasında değişmektedir.

Şekil 3.18'de AUC değerlerine göre sınıflandırıcılar şöyle yorumlanabilir: Bir numaralı sınıflandırıcı oldukça dik bir şekilde y ekseninde bir değerine yaklaşmakta ve dolayısıyla AUC değeri en yüksektir. Görece olarak bir numaralı sınıflandırıcı iki numaralı sınıflandırıcıdan, iki numaralı sınıflandırıcı üç numaralı sınıflandırıcıdan daha iyi ayrım yapabilmiştir. Dört numaralı sınıflandırıcının ise ayrım yapamadığı söylenebilir (Hoens ve Chawla, 2013).

3.4.7 PR Eğrisi ve PR-AUC Değeri

Anomali belirlemede çoğu senaryoda tüm gözlemler arasından anomalileri doğru tespit edebilmek daha önemlidir. Bu durumda PR-AUC yani kesinlik-duyarlılık grafiğinin altında kalan alanın ölçülmesi daha ayırt edici olmaktadır (Davis ve Goadrich, 2006).



Şekil 3.19 PR eğrisi

Şekil 3.19'da PR eğrisi altında kalan alan değerlerine göre 3 numaralı sınıflandırıcının diğer sınıflandırıcılardan daha başarılı olduğu söylenebilir. Tez çalışmasının uygulama sonuçları kısmında AUC değerleriyle birlikte PR-AUC sonuçları da değerlendirilmiştir.

3.4.8 F Beta Puanı

Belirleyicilik ve duyarlılık ölçütlerinin ayrı ayrı değerlendirilmesinin yeterli olmayacağı düşüncesiyle kesinlik ve duyarlılık değerlerinin harmonik ortalaması ile elde edilir. 0 ile 1 arasında çıkan değerin yüksek olması daha başarılı olduğu anlamına gelir (Japkowicz, 2013). Denklem 3.5'te hesaplanışı verilmiştir:

$$F_\beta = (1 + \beta^2)x \frac{\text{Kesinlik } x \text{ Duyarlılık}}{\beta^2 \ x \text{ Kesinlik} + \text{ Duyarlılık}} \tag{3.5}$$

Kesinlik ve duyarlılık ölçülerinin aynı öneme sahip olduğu durumda β değeri 1 alınır ve F_1 ölçüsü olarak adlandırılır. En çok kullanılan β değeri 1'dir (Japkowicz, 2013). Bu durumda Denklem 3.6'daki gibi hesaplanır.

$$F_1 = 2 \ x \ \frac{\text{Kesinlik } x \text{ Duyarlılık}}{\text{Kesinlik} + \text{ Duyarlılık}} \tag{3.6}$$

Anomali belirleme gibi yanlış pozitiflerin maliyetinin yüksek olduğu durumlarda β değeri 1'den küçük tercih edilir (Brownlee, 2021). Uygulama bölümünde β değeri 0,5 olarak alınarak F_1 sonuçlarıyla birlikte $F_{0.5}$ sonuçları da hesaplanarak incelenmiştir. Bu durumda β değeri 0,5 için Denklem 3.7'deki gibi hesaplanır.

$$F_{0.5} = (1{,}25)x \frac{\text{Kesinlik } x \text{ Duyarlılık}}{0{,}25 \ x \text{ Kesinlik} + \text{ Duyarlılık}} \tag{3.7}$$

3.4.9 Maliyet Matrisi

Toplam maliyet (TM), yanlış sınıflandırılan gözlemlerin (YP ve YN) negatif ve pozitif yanlış sınıflandırma maliyet puanları ile yanlış sınıflandırılan gözlem sayılarının çarpılması ile hesaplanır.

Maliyet Matrisi		Gerçek Durum	
		Pozitif	Negatif
Tahmin	Pozitif	0	Yanlış Pozitif Maliyeti
	Negatif	Yanlış Negatif Maliyeti	0

Şekil 3.20 Maliyet matrisi

Şekil 3.20'de matrisin sütunları gerçek değerleri göstermektedir. Matrisin satırları ise sınıflandırma neticesinde bulunan tahmin değerlerini ve yanlış sınıflandırma maliyetlerini göstermektedir.

$$TM = C(YN) \times YN + C(YP) \times YP \tag{3.8}$$

C harfi maliyet fonksiyonu olmak üzere Toplam Maliyet (TM), yanlış negatif maliyeti C(YN) ile YN gözlem sayısının çarpımı ile yanlış pozitif maliyeti C(YP) ile YP gözlem sayısının çarpımının toplanmasıyla hesaplanır. Doğru sınıflandırılan gözlemlerin bir maliyeti olmadığı için 0 maliyeti vardır. Denklemde maliyet hesabı gösterilmiştir. Burada amaç maliyeti mümkün olduğunca düşük tutmaktır. Maliyet Duyarlı Öğrenme yöntemlerinde yanılgı matrisi yerine bu matris kullanılır (Hoens ve Chawla, 2013).

BÖLÜM DÖRT
UYGULAMA

Bu bölümde, uygulanan yöntemlere, kullanılan veri kümelerine ve elde edilen sonuçlara yer verilmiştir. Uygulamada, topluluk öğrenmesi yaklaşımlarının ve yeniden örnekleme çözümlerinin, karar ağacı tabanlı denetimli anomali belirleme yöntemlerinin başarımına olan etkileri incelenmiştir. Bu çalışma kapsamında nokta anomali tespiti incelenmiştir.

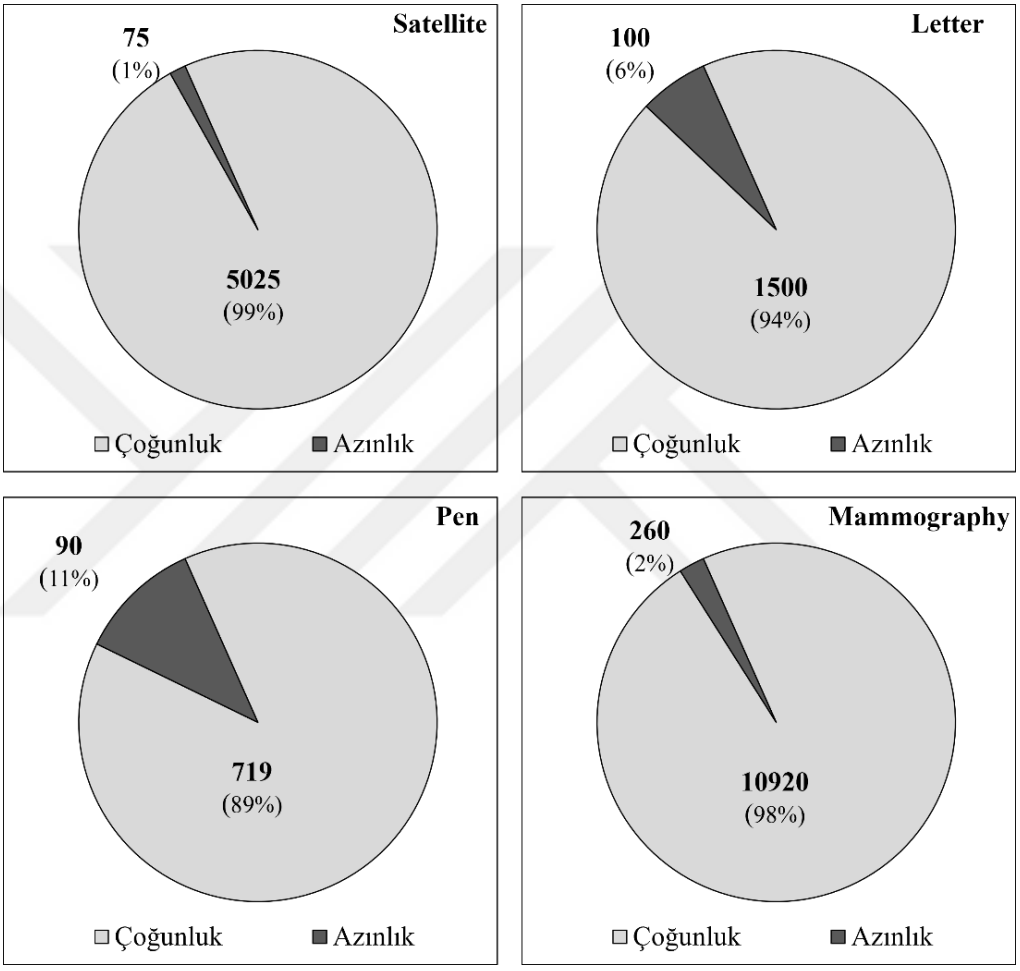
4.1 Kullanılan Veri Kümeleri ve Yöntemler

Modellerin oluşturulması ve hesaplamaların yapılması için istatistiksel hesaplama konusunda oldukça geniş kütüphanelere sahip olan açık kaynak kodlu ve yaygın kullanılan bir programlama dili olan R (R Core Team, 2022) tercih edilmiştir. Seçilen anomali belirleme yöntemleri için xgboost (Chen vd., 2022), randomForest (Liaw ve Wiener, 2002), rpart (Therneau ve Atkinson, 2022), imbalance (Cordón vd., 2018), tree (Ripley, 2021) ve ROCR (Sing vd., 2005) paketlerinden yararlanılmıştır.

Uygulamada, anomali belirleme problemi için literatürde yaygın olarak kullanılan; Unsupervised Anomaly Detection Benchmark (Goldstein, 2015) deposunda yer alan; Satellite, Letter, Pen (global) ve OpenML (Bischl, vd., 2021) deposundan indirilen Mammography (Woods vd., 1993) verileri tercih edilmiştir.

Satellite verisi uydu görüntülerinden çıkarılan özellikleri içermektedir. Uydu tarafından gözlemlenen her bölge, yeşil ve kırmızı olan görünür iki bölge ve iki kızılötesi görüntü olmak üzere dört görüntüye sahiptir. Veri, 36 bağımsız değişken, 5025 olağan ve 75 anomali olmak üzere 5100 gözlem içermektedir. Her piksel, 0'a (siyah) ve 255'e (beyaz) karşılık gelecek şekilde bir değere sahiptir. Bu piksellerin her biri 80m x 80m'lik bir alana karşılık gelmektedir. Arazi kullanımını tespit edebilmek için oluşturulan veriler içerisinden farklı toprak renkleri üzerindeki ekili kısımlar anomali olarak belirlenmiştir.

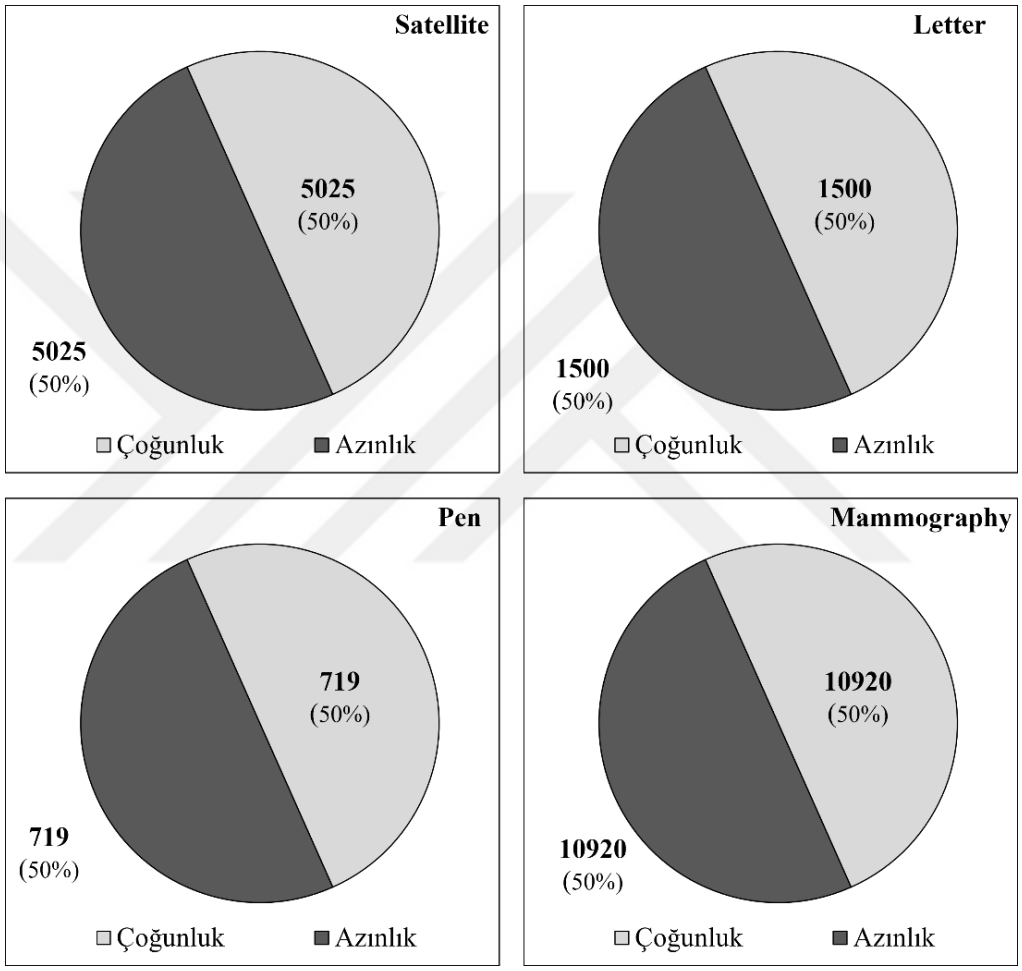
Letter verisi, İngiliz alfabesindeki 26 büyük harfin 20 farklı fontu için 16 özellik belirlenerek oluşturulmuştur. Anomali belirleme amacıyla kullanılabilmesi için orijinal veri içerisinden 3 harf olağan olarak belirlenmiş; geriye kalan harflerden alınan rastgele gözlemlerden anomaliler oluşturulmuştur (Goldstein ve Uchida, 2016). 1500 normal gözlem ve 100 anomali gözlem için 32 bağımsız değişken içermektedir.



Şekil 4.1 Kullanılan veri kümelerindeki anomali gözlem sayıları ve oranları

Pen (global) verisi 45 kişinin el yazıları ile 0'dan 9'a kadar rakamları yazmaları ile oluşturulmuştur. Her rakam için 16 bağımsız değişkene sahip özellik vektörü oluşturulmuştur. 8 rakamı korunmuş, diğer rakamlar için rassal örneklemler alınarak anomaliler eklenmiştir. 809 gözlem içermekte olup 90 gözlem anomali olarak işaretlenmiştir.

Mammography verisi, gözlemlerin değerlendirilmesine göre hastanın meme kanseri olup olmadığına dair etiket içermektedir. Veriyi oluşturan görüntüleme yöntemleri ile hücre yapılarının durumlarına göre ölçülmüş 6 adet bağımsız değişkene ait değerlerden oluşmaktadır (Woods vd., 1993). Gözlemler içerisinden meme kanseri olan hastalar anomali sınıfında değerlendirilmiştir. 11180 gözlem içerisinden 10920 hasta normal olarak, 260 hasta meme kanseri olarak etiketlenmiştir.



Şekil 4.2 Sınıf dengesizliği giderildikten sonraki oranlar

Şekilde 4.1'de veri kümesi için normal gözlem ve anomali gözlem sayıları sırasıyla çoğunluk ve azınlık sınıfı gözlemler olarak gösterilmiştir. Veri kümelerin sınıflara dağılımının dengeli olmadığı, azınlık sınıfı gözlemlerin %1 ile %6 arasında kaldığı görülmektedir.

Uygulamada ilk olarak, sınıf dengesizliği giderilmeden modeller oluşturulmuştur. Sonrasında orijinal veri kümeleri, Bölüm 3'te anlatılan yeniden örnekleme tabanlı çözümlerden rastgele aşırı örnekleme (ROS) ve SMOTE yöntemlerinden ADASYN tercih edilerek dengelenmiştir. Sınıflar arası dengesizlik giderildikten sonra modeller oluşturulmuştur. Şekil 4.2'de ROS ve ADASYN yöntemleri ile azınlık sınıfı gözlemlerin artırılmasıyla veri kümelerinin sınıf dengesizliği giderildikten sonraki sınıf dağılımları gösterilmiştir.

Her bir veri kümesinde, gözlemlerin rastgele %70'i eğitim verisi olarak ayrılarak modellerin bu veriler ile ayrı ayrı eğitilmeleri sağlanmıştır. Kalan %30'luk kısımları ise oluşturulan modellerin başarı değerlendirmeleri için test verisi olarak ayrılmıştır.

Uygulamada, topluluk öğrenmesi yaklaşımlarının ve sınıf dengesizliği gidermenin anomali belirleme performansı üzerindeki etkilerini göstermek amacıyla üç yöntem (topluluk öğrenmesi kullanılmayan Sınıflandırma Ağacı – CT, Bagging yaklaşımından Random Forest – RF, Boosting yaklaşımından XGBoost – XG) her veri için 10 kez çalıştırılmıştır. Sonuçlarda elde edilen performans ölçütlerinin ortalaması sunulmuştur. Sınıf dengesizliğini gidermenin ve topluluk öğrenmesi yöntemlerinin tercih edilmesinin daha iyi sonuçlar elde edilmesini sağlayıp sağlamadığı ampirik olarak değerlendirilmiştir.

4.2 Uygulama Sonuçları

Bu bölümde, oluşturulan modellerin başarı değerlendirme ölçütleri olarak Bölüm 3'te anlatılan anomali belirlemede tercih edilen başarı değerlendirme ölçütlerinden AUC, PR-AUC, F_1 ve $F_{0.5}$ puanları verilmiş ve yorumlanmıştır. En başarılı sonuçlar koyu yazılarak vurgulanmıştır.

4.2.1 AUC Sonuçları

Sınıf dengesizliği sorunu giderilmeden eğitilen modeller için test verileri ile elde edilen AUC değerleri Tablo 4.1'de gösterilmiştir.

Tablo 4.1 Sınıf dengesizliği sorunu giderilmeden eğitilen modeller için AUC değerleri

Veri Kümesi	Yöntem	AUC Değerleri
Satellite	CT	0,830
	RF	**0,976**
	XG	0,971
Letter	CT	0,763
	RF	**0,981**
	XG	0,969
Pen	CT	0,842
	RF	**0,999**
	XG	0,998
Mammography	CT	0,831
	RF	**0,946**
	XG	0,933

Tablo 4.1'de topluluk öğrenmesi yöntemi tercih edilmeden CT yöntemi ile anomali belirleme yapılması ile elde edilen AUC değerlerinin topluluk öğrenmesi yaklaşımları ile elde edilen sonuçlara göre tüm veri kümeleri için daha düşük olduğu görülmüştür. RF ve XG algoritmaları için AUC değerleri birbirine çok yakın olmakla birlikte RF sonuçlarının az farkla daha yüksek olduğu görülmüştür.

Sınıf dengesizliği sorunu giderilerek eğitilen modeller ile elde edilen AUC değerleri Tablo 4.2'de gösterilmiştir.

Tablo 4.2 Sınıf dengesizliği sorunu giderildikten sonra eğitilen modeller için AUC değerleri

Veri Kümesi	Yöntem	ROS	ADASYN
Satellite	CT	0,992	0,966
	RF	**1,000**	**1,000**
	XG	**1,000**	**1,000**
Letter	CT	0,941	0,863
	RF	**1,000**	**1,000**
	XG	**1,000**	0,999
Pen	CT	0,986	0,927
	RF	**1,000**	**1,000**
	XG	**1,000**	**1,000**
Mammography	CT	0,918	0,885
	RF	0,985	0,983
	XG	**0,996**	0,993

Veri dengesizliği sorunu ROS ve ADASYN yöntemleri ile giderildikten sonra eğitilen modeller üzerinde test verisi ile elde edilen AUC değerlerine göre (Tablo 4.2) tüm yöntemler için Tablo 4.1'deki sonuçlara göre iyileşme görülmüştür. CT modellere göre topluluk öğrenmesi kullanılan modellerin AUC değerleri daha yüksek çıkmıştır. ROS ve ADASYN yöntemleri kendi aralarında kıyaslandığında, ROS kullanılarak oluşturulan modellerin ADASYN ile oluşturulan modellere göre çok az farkla daha iyi sonuçlar verdiği söylenebilir.

4.2.2 PR-AUC Sonuçları

Sonuçların değerlendirilmesinde AUC değerlerinin yanı sıra incelenen problem için daha ayırt edici olduğu Bölüm 3.4.6 açıklanan PR-AUC değerleri de hesaplanmıştır. Sınıf dengesizliği sorunu giderilmeden eğitilen modeller ile elde edilen PR-AUC değerleri Tablo 4.3'te gösterilmiştir

Tablo 4.3 Sınıf dengesizliği sorunu giderilmeden eğitilen modeller için PR-AUC değerleri

Veri Kümesi	Yöntem	PR-AUC Değerleri
Satellite	CT	0,660
	RF	**0,847**
	XG	0,827
Letter	CT	0,337
	RF	**0,813**
	XG	0,812
Pen	CT	0,740
	RF	**0,996**
	XG	0,992
Mammography	CT	0,545
	RF	**0,756**
	XG	0,737

Tablo 4.3'te verilen sonuçlarda, topluluk öğrenmesi yöntemleri ile elde edilen PR-AUC değerlerinin, CT yöntemi ile elde edilen PR-AUC değerlerine göre tüm veri kümeleri için daha yüksek olduğu görülmüştür. RF yönteminin PR-AUC değerleri az farkla XG'den daha yüksektir.

Veri dengesizliği sorunu ROS ve ADASYN yöntemleri ile giderildikten sonra elde edilen PR-AUC değerlerine göre (Tablo 4.4) CT yöntemi, RF ve XG yöntemlerine göre oldukça başarısız sonuçlar vermiştir. Dengesizliğin giderilmesinin CT için tüm verilerde PR-AUC değerlerini artırmadığı görülürken RF ve XG ile elde edilen sonuçların dengesiz duruma göre iyileştiği görülmüştür. ROS ve ADASYN kıyaslandığında ROS ile oluşturulan bazı modellerin PR-AUC değerlerinin az farkla daha yüksek olduğu söylenebilir.

Tablo 4.4 Sınıf dengesizliği sorunu giderildikten sonra eğitilen modeller için PR-AUC değerleri

Veri Kümesi	Yöntem	ROS	ADASYN
Satellite	CT	0,595	0,580
	RF	**1,000**	**1,000**
	XG	0,999	0,997
Letter	CT	0,523	0,448
	RF	**1,000**	0,998
	XG	**1,000**	0,996
Pen	CT	0,870	0,790
	RF	**1,000**	**1,000**
	XG	**1,000**	**1,000**
Mammography	CT	0,300	0,204
	RF	0,**989**	0,963
	XG	0,979	0,949

Tablo 4.3 ve 4.4 sonuçlarına göre, tüm veri kümeleri için topluluk öğrenmesi yöntemleri ile sınıf dengesizliği giderildiği durumda eğitilen modellerde PR-AUC değerleri artmıştır. Ancak CT yöntemi için Letter ve Pen veri kümelerinde dengesizliğin giderildiği veri kümeleri ile eğitilen modeller üzerinde PR-AUC değerlerinde artış olduğu, Satellite ve Mammography verilerinde ise azalma olduğu görülmektedir.

Tablo 4.2 ve Tablo 4.4 birlikte değerlendirildiğinde, AUC ve PR-AUC değerleri karşılaştırıldığında, modellerin başarı sıralamalarının değişmediği ancak sınıflandırma yöntemleri arasındaki değer farklarının topluluk öğrenmesi kullanılmadığı duruma göre arttığı görülmektedir.

4.2.3 F_1 Puan Sonuçları

F_1 puanı, belirleyicilik ve duyarlılık değerleri ile model başarılarını ölçmenin tek başına yeterli olmayacağı düşüncesiyle kesinlik ve duyarlılık değerlerinin harmonik ortalaması olarak önerilmiş bir ölçüttür (Japkowicz, 2013). Veri kümelerinde sınıf

dengesizliği giderilmeden eğitilen modeller üzerinde test verisi ile elde edilen F_1 puanları Tablo 4.5'te gösterilmiştir.

Tablo 4.5 Sınıf dengesizliği sorunu giderilmeden eğitilen modeller için F_1 puanları

Veri Kümesi	Yöntem	F_1 Puanları
Satellite	CT	0,751
	RF	0,734
	XG	**0,752**
Letter	CT	0,430
	RF	0,336
	XG	**0,616**
Pen	CT	0,749
	RF	0,871
	XG	**0,884**
Mammography	CT	0,598
	RF	0,667
	XG	**0,695**

Tablo 4.5'te topluluk öğrenmesi yöntemi tercih edilmeden eğitilen modellerle elde edilen F_1 puanı sonuçlarına göre en başarılı yöntemin XG olduğu görülmektedir.

Tablo 4.6'da sınıf dengesizliği giderilerek eğitilen modeller üzerinde test verisi ile elde edilen F_1 puanları gösterilmiştir. Buna göre topluluk öğrenmesi yöntemleri ile elde edilen F_1 puanlarının CT yöntemi ile elde edilen sonuçlardan daha iyi olduğu görülmektedir. RF yöntemiyle elde edilen sonuçların da 3 veri kümesi için XG yöntemiyle elde edilen sonuçlardan daha başarılı ve bir veri kümesi için de eşit olduğu görülmektedir.

Tablo 4.6 Sınıf dengesizliği sorunu giderildikten sonra eğitilen modeller için F_1 puanları

Veri Kümesi	Yöntem	ROS	ADASYN
Satellite	CT	0,429	0,301
	RF	**0,987**	0,975
	XG	0,967	0,941
Letter	CT	0,542	0,550
	RF	**0,996**	0,977
	XG	0,973	0,967
Pen	CT	0,905	0,830
	RF	**0,998**	0,997
	XG	**0,998**	0,994
Mammography	CT	0,357	0,265
	RF	**0,958**	0,869
	XG	0,948	0,892

Tablo 4.6'da ROS ve ADASYN karşılaştırıldığında, CT ile Letter veri kümesinde elde edilen sonuç hariç tüm verilerde tüm yöntemler için ROS daha başarılıdır.

Tablo 4.5 ve Tablo 4.6 birlikte değerlendirildiğinde RF ve XG yöntemleri ile elde edilen F_1 puanlarının dengesizlik giderildiğinde arttığı görülmektedir. Ancak CT yöntemi için Satellite ve Mammography veri kümelerinde F_1 puanlarında azalma, Letter ve Pen veri kümeleri için F_1 puanlarında artış olmuştur.

4.2.4 $F_{0.5}$ Puan Sonuçları

Anomali belirlemede yanlış pozitif tahmin yapmanın maliyeti daha yüksek kabul edildiğinde, kesinliğe duyarlılıktan daha fazla önem verildiği $F_{0.5}$ puanının kullanılması önerilmektedir (Brownlee, 2021). Sınıf dengesizliği sorunu giderilmeden eğitilen modeller için test verisi ile elde edilen $F_{0.5}$ puanları Tablo 4.7'de verilmiştir.

Tablo 4.7 Sınıf dengesizliği sorunu giderilmeden eğitilen modeller için $F_{0.5}$ puanları

Veri Kümesi	Yöntem	$F_{0.5}$ Puanları
Satellite	CT	**0,846**
	RF	0,808
	XG	0,826
Letter	CT	0,478
	RF	0,441
	XG	**0,723**
Pen	CT	0,810
	RF	0,941
	XG	**0,945**
Mammography	CT	0,702
	RF	**0,774**
	XG	0,767

Tablo 4.7'deki sonuçlar incelendiğinde farklı veri kümeleri için farklı yöntemlerin başarılı olduğu görülmektedir. Satellite verisi için en yüksek $F_{0.5}$ puanı CT ile elde edilirken, Letter ve Pen verileri için XG, Mammography verisi için RF ile elde edilmiştir.

Sınıf dengesizliği giderilerek eğitilen modeller için test verisi ile elde edilen $F_{0.5}$ puanları Tablo 4.8'de gösterilmiştir. Bu senaryoda, topluluk öğrenmesi yöntemlerinin kullanılmasının $F_{0.5}$ puanına göre model başarısında oldukça etkili olduğu görülmektedir. Her iki yöntemin de $F_{0.5}$ puanını topluluk öğrenmesi yaklaşımları için olumlu, CT için olumsuz etkilediği söylenebilir. Üç veri için RF yöntemi ROS ile, bir veri için yine RF yöntemi ADASYN ile daha başarılı sonuçlar vermiştir.

Tablo 4.8 Sınıf dengesizliği sorunu giderildikten sonra eğitilen modeller için $F_{0.5}$ puanları

Veri Kümesi	Yöntem	ROS	ADASYN
Satellite	CT	0,322	0,215
	RF	**0,980**	0,964
	XG	0,950	0,912
Letter	CT	0,400	0,437
	RF	**0,988**	0,969
	XG	0,951	0,969
Pen	CT	0,862	0,817
	RF	0,997	**0,999**
	XG	0,997	0,998
Mammography	CT	0,264	0,188
	RF	**0,946**	0,820
	XG	0,931	0,865

$F_{0.5}$ puan sonuçları birlikte değerlendirildiğinde, dengesizliği gidermenin CT yöntemi için önemli bir iyileştirme sağlamadığı ancak topluluk öğrenmesi yaklaşımının kullanıldığı yöntemlerde tüm veriler için daha yüksek $F_{0.5}$ puanı elde edilmesini sağladığı görülmüştür.

BÖLÜM BEŞ
DEĞERLENDİRME VE SONUÇ

Bu tez çalışmasında topluluk öğrenmesi yöntemleriyle anomali belirleme incelenmiştir. Bu konuda daha önce yapılan ilgili çalışmalar araştırılmış; anomali belirleme, topluluk öğrenmesi ve sınıf dengesizliğini gidermede yeniden örnekleme yaklaşımlarına yer verilmiştir.

Tez çalışmasında anomali belirleme bir azınlık sınıfı belirleme problemi olarak ele alınmış ve iki temel soruya yanıt aranmıştır. Birinci soru topluluk öğrenmesi yaklaşımlarının kullanılmasının, anomali belirleme başarımına etkisinin olup olmadığıdır. İkinci soru ise veri kümesindeki sınıf dengesizliğinin giderilmesinin sınıflama başarımına katkı sağlayıp sağlamadığıdır.

Bu kapsamda, geleneksel sınıflandırma ağacı-CT, Bagging yaklaşımından Random Forest-RF, Boosting yaklaşımından XGBoost-XG yöntemleri, rastgele aşırı örnekleme-ROS ve ADASYN yeniden örnekleme yöntemleri, anomali oranları %1 ile %6 arasında değişen dört farklı anomali veri kümesi (Satellite, Letter, Pen-global ve Mammography) üzerinde uygulanmıştır. Uygulamada hem sınıf dengesizliği giderilmeden hem de ROS ve ADASYN yöntemleri ile yeniden örnekleme yapılarak sınıf dengesizliği giderilmiş veride CT, RF ve XG yöntemleri ile modeller oluşturulmuştur. Oluşturulan modellerin performans ölçütleri olarak AUC ve PR-AUC değerleri ile F_1 ve $F_{0.5}$ puanları kullanılmıştır.

CT yöntemi ile elde edilen AUC ve AUC-PR değerlerinin hem dengesiz hem de dengesizliğin giderildiği senaryolarda topluluk öğrenmesi yaklaşımları ile elde edilenlerden düşük olduğu görülmüştür. RF ve XG algoritmaları için tüm senaryolarda değerler birbirine çok yakın olmakla birlikte RF sonuçlarının az farkla daha yüksek olduğu söylenebilir. Sınıf dengesizliğini gidermenin topluluk öğrenmesi yöntemlerinin başarısını tüm veri kümelerinde artırdığı ancak topluluk öğrenmesi kullanılmadığında sonuçların veriye göre değişiklik gösterdiği görülmüştür.

F_1 ve $F_{0.5}$ puanları incelendiğinde, CT yöntemi için sınıf dengesizliğini gidermenin her veri kümesinde performansı artırmadığı görülmüştür. RF ve XG yöntemleri için tüm durumlarda puanlar birbirine çok yakın gözlenmiştir. F_1 ve $F_{0.5}$ puanları açısından da dengesizliği gidermenin topluluk öğrenmesi yaklaşımının kullanıldığı yöntemler için önemli iyileştirme sağladığı görülmüştür. F_1 ve $F_{0.5}$ puanlarına göre en iyi sonuçlar RF'nin ROS ile birlikte kullanımda elde edilmiştir.

Sonuçlarda genel olarak ortaya çıkan topluluk öğrenmesi kullanılmayacak ise sınıf dengesizliğini gidermenin çok etkili olmadığı, topluluk öğrenmesi yaklaşımları kullanılacaksa dengesizliği gidermenin incelenen tüm performans ölçütleri için etkili olduğudur.

Sonuçlar genel olarak değerlendirildiğinde, sınıf dengesizliğini gideren yöntemler ile birlikte kullanıldığında topluluk öğrenmesi yaklaşımlarının anomali belirleme (azınlık sınıfı belirleme) problemlerinde daha başarılı sonuç verdiği görülmüştür. Ayrıca, bu problem için topluluk öğrenmesi kullanılmayacak ise sınıf dengesizliğini gidermenin çok etkili olmadığı da söylenebilir. Sonraki çalışmalarda bu sonuçların diğer topluluk öğrenmesi yaklaşımları için de geçerli olup olmadığının araştırılması konu edilebilir.

KAYNAKLAR

Aggarwal, C. C. ve Sathe, S. (2017). *Outlier ensembles*. Springer International Publishing.

Aggarwal, C. C. (2017). *Outlier analysis*. Springer International Publishing AG. https://doi.org/10.1007/978-3-319-47578-3

Baesens, B., Vlasselaer, V., V. ve Verbeke, W. (2015). *Fraud analytics using descriptive, predictive, and social network techniques.* John Wiley & Sons.

Batista, G. E. A. P. A., Prati, R. C. ve Monard, M.C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, 6(1), 20-29.

Bischl, B., Casalicchio, G., Feurer, M., Gijsbers, P., Hutter, F., Lang, M., Mantovani, R., G., Rijn, J. N. ve Vanschoren J. (2021). *OpenML: A benchmarking layer on top of OpenML to quickly create, download, and share systematic benchmarks.* https://www.openml.org/search?type=data&status=active&id=310&sort=runs

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2), 123-140.

Brownlee, J. (20 Haziran 2022), *A gentle introduction to the fbeta-measure for machine learning*. https://machinelearningmastery.com/fbeta-measure-for-machine-learning

Chandola, V., Banerjee, A. ve Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, 41(3), 1-58.

Chawla, N. V., Bowyer, K. W., Hall, L. O. ve Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique, *Journal of artificial intelligence research*, 16, 321-357.

Chawla, N.V., Lazarevic, A., Hall, L.O. ve Bowyer, K.W. (2003). SMOTEBoost: Improving prediction of the minority class in boosting. Lavrač, N., Gamberger, D., Todorovski, L., Blockeel, H. (Ed). *European conference on principles of data mining and knowledge discovery* içinde (107-119). Springer, Berlin, Heidelberg.

Chawla, N.V. ve Sylvester, J. (2007). Exploiting diversity in ensembles: Improving the performance on unbalanced datasets. Haindl, M., Kittler, J., Roli, F. (Ed). *Multiple Classifier Systems* içinde (397-406). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-72523-7_40

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., Li, Y. ve Yuan, J. (2022). *xgboost: Extreme gradient boosting. R paketi versiyon 1.6.0.1*. https://CRAN.R-project.org/package=xgboost

Chen, T. ve Guestrin, C. (2016). *Xgboost: A scalable tree boosting system*. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. New York, NY: ACM.

Cordón, I., García, S., Fernández, A. ve Herrera, F. (2018). Imbalance: Oversampling algorithms for imbalanced classification in R. *Knowledge-Based Systems, 161*, 329-341.

Das, B., Krishnan, N. C. ve Cook, D. J. (2014). RACOG and wRACOG: Two probabilistic oversampling techniques. *IEEE transactions on knowledge and data engineering*, *27*(1), 222-234.

Davis, J. ve Goadrich, M. (2006). *The relationship between Precision-Recall and ROC curves*. 23rd international conference on Machine learning, 233-240.

Drummond, C. ve Holte, R. C. (2003). C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. *Workshop on learning from imbalanced datasets II* (11), 1-8.

Estabrooks, A., Jo, T. ve Japkowicz, N. (2004). A multiple resampling method for learning from imbalanced data sets. *Computational intelligence*, *20*(1), 18-36.

Efron, B. (1992). Bootstrap methods: another look at the jackknife. *Breakthroughs in statistics*, (569-593). Springer, New York, NY.

Fernandes, E. R. ve Carvalho, A. C. (2019). Evolutionary inversion of class distribution in overlapping areas for multi-class imbalanced learning. *Information Sciences*, 494, 141-154.

Fernandez, A., Garcia, S., Galar, M., Prati, R. C., Krawezyk, B. ve Herrera, F. (2018). *Learning from imbalanced data sets*. Berlin: Springer.

Freund, Y. ve Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, *55*(1), 119-139.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of statistics*, 29(5), 1189-1232.

Geurts, P., Ernst, D. ve Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63(1), 3-42. https://doi.org/10.1007/s10994-006-6226-1

Goldstein, M. Ve Uchida, S. (2016). A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PloS one*, *11*(4), e0152173.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H. ve Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert systems with applications*, 73, 220-239.

Hanifah, F. S., Wijayanto, H. ve Kurnia, A. (2015). Smotebagging algorithm for imbalanced dataset in logistic regression analysis (Case: Credit of bank x). *Applied Mathematical Sciences*, 9(138), 6857-6865.

Hastie, T., Tibshirani, R. ve Friedman, J. (2009). Random forests. *The elements of statistical learning* içinde (587-604). Springer, New York, NY.

He, H., Bai, Y., Garcia, E. A. ve Li, S. (2008). *ADASYN: Adaptive synthetic sampling approach for imbalanced learning*. 2008 IEEE international joint conference on neural networks, 1322-1328

He, H. ve Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9), 1263-1284.

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, 20(8), 832-844.

Hoens, T. R. ve Chawla, N. V. (2013). Imbalanced datasets: from sampling to classifiers. *Imbalanced learning: Foundations, algorithms, and applications*, 43-59.

Japkowicz, N. (2013). Assessment metrics for imbalanced learning. *Imbalanced learning: Foundations, algorithms, and applications*, 187-206.

Kearns, M. ve Valiant, L. (1994). Cryptographic limitations on learning boolean formulae and finite automata. *Journal of the ACM (JACM)*, 41(1), 67-95.

Keogh, E., Lonardi, S. ve Chiu, B. Y. C. (2002). Finding surprising patterns in a time series database in linear time and space. *Eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 550-556.

Krawczyk, B. (2016). Learning from imbalanced data: Open challenges and future directions. *Progress in Artificial Intelligence*, 5(4), 221-232.

Kuhn, M. ve Johnson, K. (2013). *Classification trees and rule-based models*. Applied predictive modeling (369 - 413). Springer, New York, NY. https://doi.org/10.1007/978-1-4614-6849-3_14

Liaw, A. ve Wiener, M. (2002). Classification and regression by randomForest. *R news*, *2*(3), 18-22.

Liu, F. T., Ting, K. M. ve Zhou, Z. H. (2008). Isolation forest. *Eighth IEEE international conference on data mining*, (413-422). IEEE.

Liu, X. Y., Wu, J. ve Zhou, Z. H. (2008). Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2), 539-550.

Menardi, G. ve Torelli, N. (2014). Training and assessing classification rules with imbalanced data. *Data mining and knowledge discovery*, *28*(1), 92-122.

Prabhakaran, S. (20 Mayıs 2022). *Outlier detection and treatment with R*. https://datascienceplus.com/outlier-detection-and-treatment-with-r.

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Ripley, B. (2021). tree: Classification and regression trees. R paketi versiyon 1.0-41, https://CRAN.R-project.org/package=tree.

Schapire, R. E. (1990). The strength of weak learnability. *Machine learning*, 5(2), 197-227.

Sagi, O. ve Rokach, L. (2018). Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(4), e1249.

Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J. ve Napolitano, A. (2008). RUSBoost: Improving classification performance when training data is skewed. *2008 19th international conference on pattern recognition*, (1-4). IEEE.

Sing, T., Sander, O., Beerenwinkel, N. ve Lengauer, T. (2005). ROCR: Visualizing classifier performance in R. *Bioinformatics*, *21*(20), 3940-3941.

Sun, J., Lang, J., Fujita, H. ve Li, H. (2018). Imbalanced enterprise credit evaluation with DTE-SBD: Decision tree ensemble based on SMOTE and bagging with differentiated sampling rates. *Information Sciences*, *425*, 76-91.

Sutton, C. D. (2005). Classification and regression trees, bagging, and boosting. *Handbook of statistics*, 24, 303-329.

Therneau, T. ve Atkinson, B. (2022). rpart: Recursive partitioning and regression trees, R paketi versiyon 4.1-16. https://CRAN.R-project.org/package=rpart.

Wang, S. ve Yao, X. (2009). *Diversity analysis on imbalanced data sets by using ensemble models*. 2009 IEEE symposium on computational intelligence and data mining (324-331). IEEE.

Wang, H., Bah, M. J. ve Hammad, M. (2019). Progress in outlier detection techniques: A survey. *IEEE Access*, *7*, 107964-108000.

Wolpert, D. H. (1992). Stacked generalization. *Neural networks*, 5(2), 241-259.

Woods, K. S., Doss, C. C., Bowyer, K. W., Solka, J. L., Priebe, C. E. ve Kegelmeyer Jr, W. P. (1993). Comparative evaluation of pattern recognition techniques for detection of microcalcifications in mammography. *International Journal of Pattern Recognition and Artificial Intelligence, 7*(06), 1417-1436.

Yıldıztepe, E. ve Akıllı, N. (2021) *Dengesiz verilerde boosting ile anomali belirleme üzerine bir çalışma* [Bildiri Özeti]. 4th International Conference on Data Science and Applications, Muğla, Türkiye. http://www.icondata.org/uploads/kcfinder/upload/files/ICONDATA21_Full_Text_Book_final.pdf

Polikar, R. (2012). Ensemble learning. *Ensemble machine learning* (1-34). Springer, Boston, MA.

Zhou, Z. H. (2012). *Ensemble methods: foundations and algorithms* (1. Baskı). Chapman & Hall/CRC. https://dl.acm.org/doi/book/10.5555/2381019

Zhou, Z. H. Ve Liu, X. Y. (2005). Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on knowledge and data engineering*, *18*(1), 63-77.